Quantum Cognitively Motivated Context-Aware Multimodal Representation Learning for Human Language Analysis

Thesis

QUANTUM COGNITIVELY MOTIVATED CONTEXT-AWARE MULTIMODAL REPRESENTATION LEARNING FOR HUMAN LANGUAGE ANALYSIS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF COMPUTING AND COMMUNICATIONS
AND THE COMMITTEE ON GRADUATE STUDIES
OF THE OPEN UNIVERSITY (UK)
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Dimitris Gkoumas
May 2021

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Dawei Song) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Yijun Yu)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Shailey Minocha)

Approved for the Open University (UK) Committee on Graduate Studies

Abstract

A long-standing goal in the field of Artificial Intelligence (AI) is to develop systems that can perceive and understand human multimodal language. This requires both the consideration of context in the form of surrounding utterances in a conversation, i.e., *context modelling*, as well as the impact of different modalities (e.g., linguistic, visual acoustic), i.e., *multimodal fusion*. In the last few years, significant strides have been made towards the interpretation of human language due to simultaneous advancement in deep learning, data gathering and computing infrastructure. AI models have been investigated to either model interactions across distinct modalities, i.e., linguistic, visual and acoustic, or model interactions across parties in a conversation, achieving unprecedented levels of performance. However, AI models are often designed with only performance as their design target, leaving aside other essential factors such as transparency, interpretability, and how humans understand and reason about cognitive states.

In line with this observation, in this dissertation, we develop quantum probabilistic neural models and techniques that allow us to capture rational and irrational cognitive biases, without requiring *a priori* understanding and identification of them. First, we present a comprehensive empirical comparison of state-of-the-art (SOTA) modality fusion strategies for video sentiment analysis. The findings provide us helpful insights into the development of more effective modality fusion models incorporating quantum-inspired components. Second, we introduce an end-to-end complex-valued neural model for video sentiment analysis, simulating quantum procedural steps, outside of physics, into the neural network modelling paradigm. Third, we investigate non-classical correlations across different modalities. In particular, we describe a methodology to model interactions between image and text for an information retrieval scenario. The results provide us with theoretical and empirical insights to develop a transparent end-to-end probabilistic neural model for video emotion detection in conversations, capturing non-classical correlations across distinct modalities. Fourth, we introduce a theoretical framework to model user's cognitive states underlying their multimodal decision perspectives, and propose a methodology to capture interference of modalities in decision making.

Overall, we show that our models advance the SOTA on various affective analysis tasks, achieve high transparency due to the mapping to quantum physics meanings, and improve post-hoc interpretability, unearthing useful and explainable knowledge about cross-modal interactions.

Acknowledgements

There are many people I must thank for contributing to the three wonderful years of my experience as a PhD student.

First, I must thank my principal adviser Professor Dawei Song who chiselled me from an eager but mostly confused student to a competent researcher. When I pitched an idea, he encouraged me and provided me freedom to develop it, distilled its essence, uncovered the more fundamental story and placed it in a broader context. When I produced terrible paper drafts, he patiently provided me help until I got it. I am very grateful to Dawei Song for teaching me not just about some aspects of artificial intelligence or the art to communicate ideas, but for teaching me how to think.

I would also like to thank my supervisors Dr. Yijun Yu and Professor Shailey Minocha for their help and inputs whenever I needed them and also for the great discussions with them on research related things. I am also thankful to Professor Arosha Bandara for his general help and support.

During my PhD, I also squeezed three three-month secondments in Italy, Denmark, and Canada, and a one-month research visit in Tianjin/China. All of them were a wonderful learning experience that had a lot of impact on my research trajectory. From these secondments, I would especially like to thank Professor Massimo Melucci from the University of Padua, Professor Christina Amalia Lioma from the University of Copenhagen, and Professor Jian Yun Nie from the University of Montreal.

I would like to thank my close collaborators who I've had the distinct pleasure of working with and learning from. I would like to especially thank Qiuchi Li, who is a seemingly infinite generative model of unique perspectives and insightful comments, feedback and advice. My thinking and research philosophy has similarly been shaped by thoughtful discussions and interactions with him. I would also like to thank Sagar Uprety, who has all triggered my thoughts on innovation and who inspired me to think bigger and aim higher. It was also a great pleasure to collaborate with Dr Shahram Dehdashti. I would also like to thank Prayag Tiwari for his support and talking about things research and beyond. I am also grateful to QUARTZ project consortium for the many project meetings all over Europe and all of my colleagues.

I am thankful to Quantum Information Access and Retrieval Theory (QUARTZ) project, which has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 721321, that supported my research. I am

also thankful to the Open University's School of Computing and Communications that extended my funding for finalizing my PhD thesis.

A big thanks to those little angels of my day-to-day life who helped me to overcome obstacles and believe in me ten times more than I do.

Acronyms

AI Artificial Intelligence. iv, 1, 2, 15, 17, 124–128

CNN Convolutional Neural Network. 16

IR Information Retrieval. xvi, 9–12, 43, 51, 71, 72, 75, 78

LSTM Long Short-term Memory. 16

NLP Natural Language Processing. 6, 16, 41, 43, 81

POVM Positive Operator-valued Measure. 7, 46

PVM Projected-Valued Measure. 45, 46

QT Quantum Theory. 7, 10, 11, 42, 43, 49, 51, 70, 72, 79, 81, 91, 108, 109, 126–128

RNN Recurrent Neural Network. 5, 16, 17

SOTA state-of-the-art. iv, viii, xii, xv, 5, 6, 9–11, 15, 17, 18, 20, 21, 24–28, 32–35, 38–40, 42, 49, 50, 61, 69, 81, 82, 88, 90, 96, 99, 101, 105–107, 122, 123, 125, 127

XAI eXplainable AI. 51

Contents

List of Tables

List of Figures

Chapter 1

Introduction

One of the ultimate goals of AI is to enable computers to understand human language and endow them with the ability to communicate with us, in the multiparty setting, producing replies in natural language that are tuned to a specific context.

Human language is inherently *multimodal* and is manifested via words (i.e., linguistic modality), gestures (i.e., visual modality), and vocal intonations (i.e., acoustic modality). In general terms, in computer science, *modality* refers to a certain type of information and/or the representation format in which information is stored [13]. The means whereby information is delivered to the senses of an interpreter is called medium. Moreover, a research problem is characterised as multimodal when it includes multiple such modalities. On the other hand, human language is *contextual* usually in the form of preceding utterances. Following the above inspiring vision in AI, the scope of this dissertation is to investigate and deploy methods to model interactions across different modalities and context of sequential preceding utterances for analysing human language.

1.1 The "Holy Grail" of Understanding Human Language

Human language understanding task Giving machines the capability to understand human language effectively opens new horizons for human-machine conversation systems [148], tutoring systems [107], and health care [117], to name a few applications. Even though for humans, comprehending human language is an effortless task, this is a non-trivial challenge for machines. For computers, understanding human language requires both the consideration of *context modelling* in the form of surrounding utterances in a conversation, and *multimodal fusion* to modelling interactions of different modalities that constitute human language into a compact multimodal representation. As far as context modelling concerning, AI models should consider speaker information of utterances and the relative position of other utterances from the target utterance. Speaker information is about

modelling *inter-speaker* dependency, which aids an AI model to understand how a speaker influences other speakers, e.g., to change an opinion. Similarly, *self-dependency* enables an AI model to understand how a speaker resists, e.g., to the change of his mind, against external influence. On the other hand, consideration of relative position of target and preceding utterances decides how past utterances influence the current utterance. As far as multimodal fusion concerning, it depends on modelling each modality individually in a way that can be interacted with other modalities, a.k.a., *intra-modal dynamics*, and modelling interactions across different modalities, a.k.a., *inter-modal* or *cross-modal dynamics*.

Multimodal intelligence The central challenge of a multimodal problem is to model inter-dependencies and complementary presence in heterogeneous data originating from multiple modalities, i.e., inter-modal dynamics, that change the overall perception of a real-world problem. For instance, let us consider the situation in which we want to identify the emotional state of the utterance "Oh yeah, that's right!". In the absence of other modalities, the utterance itself is ambiguous, since there is no specific context to bias the emotional state of the utterance. For example, the utterance might express emotions of anger, sadness, joy, or surprise (see Figure 1.1). If we consider that the person speaks loudly "Oh yeah, that's right!", its emotional state could express positive emotions (i.e., joy or surprise) or a negative emotion (i.e., anger), excluding sadness which is commonly expressed by soft voice (see Figure 1.1). Even though *bi-modal* interactions make less ambiguous the utterance, its emotional state can not be safely clarified. However, as soon as all modality behaviours are simultaneously present, the utterance's emotional state becomes apparent (see tri-modal in Figure 1.1). This implies that distinct modalities should not be considered in isolation; rather, each modality acts as a context for the other modalities. The fusion of distinct modalities, a.k.a., *multimodal fusion*, is hence contextual, and this is a subtle issue.

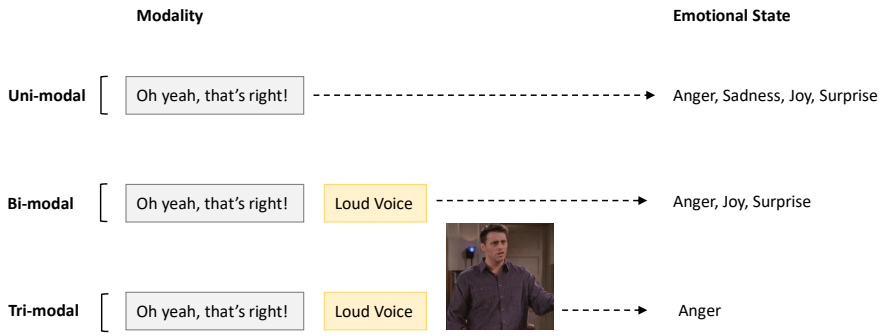


Figure 1.1: The above figure shows how uni-modal, bi-modal, and tri-modal dynamics disambiguate the emotional state of a multimodal utterance.

In contrast to verbal information (i.e., linguistic) where some kind of structure is provided by its elements such as punctuation and paragraphs, non-verbal information (i.e., visual, acoustic) is

typically an uninterrupted stream with few delimiters. For non-verbal information, defining the semantic unit is a fundamental step to attain high-quality representation. In other words, there is often a large gap between the content of a multimedia signal and its meaning. This is usually referred to as the *semantic gap*. While there is a semantic gap between the words in a sentence and its overall information and meaning, there is an even larger gap between non-verbal information and its semantics, and the gap gets bigger and bigger as the type of signal increases in complexity [12] (see Figure 1.2). On the other hand, interactions across different modality streams can be *synchronous* or

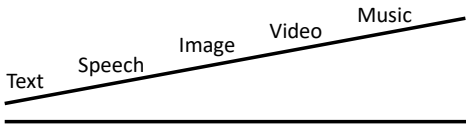


Figure 1.2: The semantic gap between multimedia streams and their meaning gets larger as signals increase in complexity.

asynchronous. A multimodal interaction is characterized as synchronous when it occurs at the same timestamp. For instance, a simultaneous co-occurrence of a smile with a positive word (see Figure 1.3, Top). By contrast, a multimodal interaction is characterized as asynchronous when spans over a long-range multimodal sequence. For instance, a prolonged occurrence of laughter with a positive word (see Figure 1.3, Bottom). Alignment strategies, such as attention mechanisms, is a common practice to deal with both synchronous and asynchronous cross-modal interactions [126].

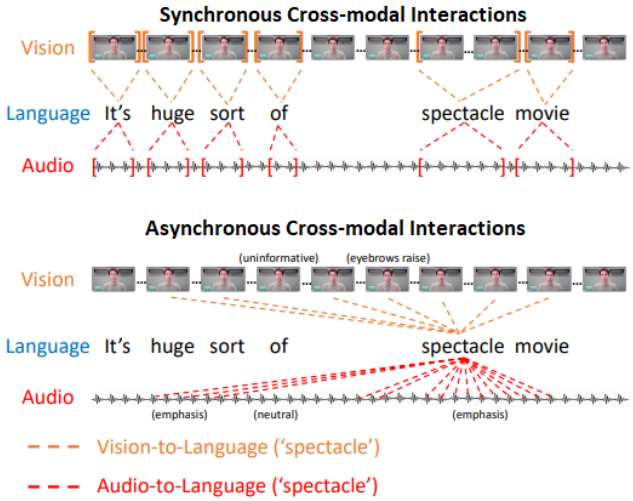


Figure 1.3: Example video clip from movie reviews. [Top]: Illustration of word-level alignment where video and audio features are averaged across the time interval of each spoken word, i.e., synchronous cross-modal interactions. [Bottom] Illustration of crossmodal attention weights between text ("spectacle") and vision/audio, i.e., asynchronous cross-modal interactions.

Representation learning Currently, deep learning, as a particular area within representation learning, copes with the semantic gap within modalities. In particular, it uses artificial neural networks with many hidden layers transforming raw data to higher-level representations or features for a specific task [16]. Representation learning is of a great value in practice since better representations can often simplify learning tasks [37, 69, 118]. By contrast, *multimodal representation learning* is not only constrained to transform the raw data to higher-level representations, i.e., modelling of intra-modality dynamics. Additionally, it primarily aims at fusing different modality signals into a compact multimedia representation and thereby enables cross-modality signal processing. Multimodal representations can be categorized into *joint* and *coordinated* [13]. Joint representations are projected to the same space using all of the modalities as input. Coordinated representations, on the other hand, exist in their own space, but are coordinated through a similarity (e.g. based on Euclidean distance) or structure constraint (e.g. partial order). In summary, representation learning deals with the semantic gap of multimedia streams, while multimodal representation learning handles synchronous and asynchronous cross-modal interactions.

Information fusion *Fusion*, being a key research topic in multimodal problems, integrates information extracted from different unimodal data sources into a single compact multimodal representation. Fusion methods can be divided based on the stage in which fusion occurs during the associated procedures (see Figure 1.4). *Early fusion*, a.k.a., *feature-level or content-level fusion*, integrates features after being extracted. In this dissertation, early fusion, content-level, and feature-level fusion are interchangeable. Feature-level fusion is mainly advantageous in that it can accommodate cross-modal interactions at an early stage. However, features should be represented in the same format before fusion. *Late fusion*, a.k.a., *decision-level fusion*, which aggregates decisions of separated models for each modality, has been regarded as more flexible, as it does not require the costly cross-feature synchronisation. However, it lacks feature-level interactions across different modalities. Again, late fusion and decision-level fusion are interchangeable. Overall, feature-level and decision-level fusions can suppress either intra- or inter-modality interactions, respectively [159]. However, since deep learning essentially involves learning hierarchical representations from raw data, this gives rise to a more flexible multimodal fusion approach, called *intermediate-level fusion*, where a shared representation layer is constructed by merging units with connections coming into this layer from multiple modality-specific paths (see Figure 1.4 (C)). That is, representations are learned using different kinds of neural layers and fused using fusion layers, also known as a *shared representation layers* [119].

In this dissertation, we consider verbal (i.e., linguistic) and non-verbal (i.e., visual and acoustic) modalities, exploit deep learning representations to model intra-modality dynamics, and leverage feature-level, intermediate-level, and decision-level modality fusion strategies to deploy joint computational models for human language analysis.

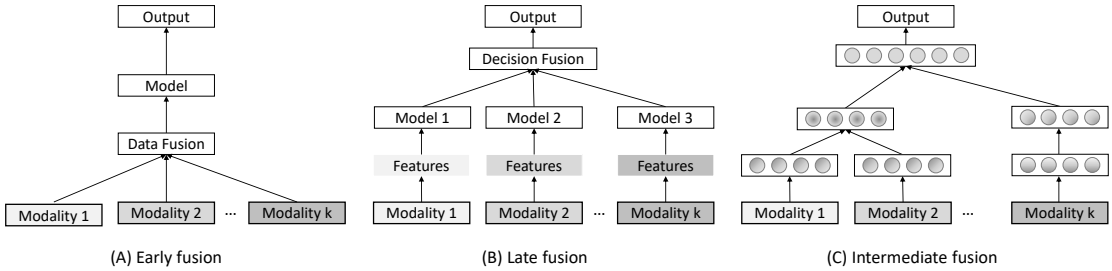


Figure 1.4: Fusion methods based on the stage in which modality fusion occurs.

1.2 Encouraging Progress and Remaining Challenges

Encouraging progress. Despite the complexity of the above-described tasks, in recent years, research has made significant strides towards the understanding of human language, and especially in the field of affective computing, due to simultaneous advances in state-of-the-art (SOTA) benchmarking datasets for video sentiment analysis and emotion recognition [27, 115, 155, 157]. Due to recent advances in deep learning, a recent trend was geared towards Recurrent Neural Network (RNN) [95] backbone architectures, which treats human language analysis task as either multimodal fusion or context modelling problem. In particular, neural approaches have been investigated to either model interactions across distinct modalities, i.e., linguistic, visual, and acoustic, [115, 126, 143, 152, 151], or model interactions across parties in a conversation [48, 59, 60, 92], after merging different modalities into a joint multimodal representation.

Remaining challenges. Recent advances in affective computing are gearing towards complex Artificial Intelligence (AI) models, achieving unprecedented performance. However, most of the current AI models cannot overcome the "black-box problem", leading to limited explainability in real-life applications. To tackle this problem, currently, researchers shall dwell in the field of eXplainable AI (XAI). The field of XAI aims to equip AI learning models with transparency, fairness, accountability and explainability [9]. To this end, AI models should adapt concepts from cognitive sciences, which provide methodological tools to explain the decisions made. In particular, a step forward beyond the high performance, when deploying AI models, should be the consideration of how humans understand and reason about the cognitive states behind the decisions. First, the exploitation of a cognitive framework in the deployment of an AI model could mitigate cognitive biases while taking decisions. Besides, it could draw attention to the various adversarial perturbations, which would lead to wrong predictions or decisions. More importantly, it could give an insight into the causality established by the learning model as well as the reasoning of the model. Hence to use AI models for understanding human language in real life, we need to make AI models accountable by explaining their decisions and make them transparent, forming the building blocks for Responsible or Ethical AI. Moreover, the new alternative design factors should not sacrifice the

high performance that existing neural models achieve in affective computing. That is, there is a challenge for researchers to distil cognitive biases into workable computational models.

1.3 Motivation of Using Quantum Theory to Model Human Language and Challenges

Paradoxes in decision making. Although SOTA neural network approaches have demonstrated an excellent performance, they neglect how people understand and reason about cognitive states [102]. However, it is widely known in the literature of cognitive science that when it comes to decision making under uncertainty, humans usually choose preferences that do not always obey the classical (Kolmogorov) axioms of probability [68] and utility theory [97], leading to decisions that are either sub-optimal, paradoxical or even irrational [4, 66, 129]. Thus, models that could take into account not only rational decisions but also irrational and paradoxical human decisions could greatly benefit decision support systems.

Quantum Cognition One general framework that has been extensively studied in the domain of human cognition and decision making is quantum probability theory [25]. The framework accommodates not only optimal rational decisions, but also several sub-optimal, irrational, and paradoxical decisions reported in the literature [128], without requiring *apriori* understanding and identification of cognitive biases. Conceptually, Quantum Theory (QT) challenges the notion that the outcome of a measurement is merely extracting information about "pre-existing" elements of reality [40]. Conversely, QT invalidates this assumption, showing that reality does not always consist of object behaving objectively [158]. Generally, in QT, the outcome of measurement will depend on other observations that are made. Similar to multimodal fusion, QT is also contextual.

QT is not only a physical theory but also a framework in which theories can be developed. Indeed, QT has increasingly been deployed outside physics. Early work showed that in some cases human language understanding exhibits certain non-classical phenomena, such as "entanglement" [18], i.e., non-classical correlations, "quantum interference" [140], and ambiguities [19], enabling quantum probabilities to serve as a suitable framework for modelling human language. Recently, the quantum measurement postulate, in conjunction with a set of procedural steps to transform a classical model to its quantum analogue, has been simulated into the neural network modelling paradigm for Natural Language Processing (NLP) tasks [78, 139]. The quantum-probability networks have not only achieved a SOTA performance but also demonstrated a high-level explainability in terms of model transparency, due to their theoretical roots in the well-established quantum physics meanings. A more detailed description of related work in this direction will be presented in the next chapters. Nevertheless, the investigation of the mathematical framework of quantum mechanics in modelling human language is preliminary, focusing on small-scale user studies, and considers only linguistic modality. It is an open research question of how such a mathematical framework of high

theoretical significance could be applied to the multimodal setting, tested on larger benchmarking datasets, and yield a practical significance.

Short-term and long-term motivations. Fusing modalities with a quantum-driven way can be motivated from both a concrete (short-term) perspective and a general (long-term) perspective.

In terms of short-term concrete motivations, first, quantum-inspired representation learning exhibits a better model transparency due to its theoretical roots in the well-established quantum physics meanings. Additionally, in this dissertation, we introduce strategies which are in line with standard procedural steps to convert information to its quantum analogue, namely, *preparation*, *evolution*, and *measurement*. In fact, due to the complexity of models, their interpretability decreases while the performance increases. Second, the mathematical formalism of quantum theory allows distilling cognitive biases into neural workable models and strategies, modelling humans' understanding and reasoning of cognitive states. Third, due to inherent properties of quantum mechanics, quantum-inspired multimodal representation learning is capable of optimizing post-hoc interpretability, unearthing useful and explainable knowledge about the way distinct modalities interact with each other and contribute to the final decision made.

In terms of long-term general motivations, the QT mathematical framework, capturing classical and irrational cognitive biases, could be leveraged for multiparty task-oriented dialogue systems, adapting the agents' retrieve content to the context of a conversation [144], and improve dialogue components which manipulate a conversation[73].

Challenges of this approach. A crucial challenge of mapping information to its quantum analogue is that information is represented as a state vector of unit length in a basis of mutually orthogonal vectors. Despite there exist algorithms for training mutually orthogonal vectors [8, 147], such a constraint makes computational models not affordable. In this dissertation, we overcome such a limitation, by training bases into the neural network modelling paradigm that their basis vectors do not necessarily form an orthonormal basis. From an interpretation point of view, basis vectors correspond to abstract semantic concepts, which are not necessarily independent of each other in practice. From a quantum point of view, we exploit Positive Operator-valued Measure (POVM) [99], which are not necessarily orthogonal. Although the impact of orthogonality in quantum probabilistic models is beyond the scope of the current dissertation, a unitary algorithm has been preliminary investigated in a joint work to address the evolution of utterances in a conversation with a quantum view [76].

1.4 Research Questions

This dissertation aims at developing transparent and unified quantum probabilistic computational models for human language analysis, endowing the quantum-cognitive nature of the human decision-making process, with a focus on rational and paradoxical cognitive biases. To this end, we borrow

concepts from QT to endow computational models with quantum cognition. Hence, the main research question of this dissertation is:

Main RQ If and how can a mathematical framework of Quantum Theory be applied into workable computational models to accommodate rational and irrational cognitive biases, underlying the multimodal decision perspectives, and yield a practical significance for analyzing human language?

The answer to this question is sought by breaking it down into the four sub-research questions, which are described below. Figure 1.5 shows how these research questions are investigated and answered in the specific chapters of the thesis.

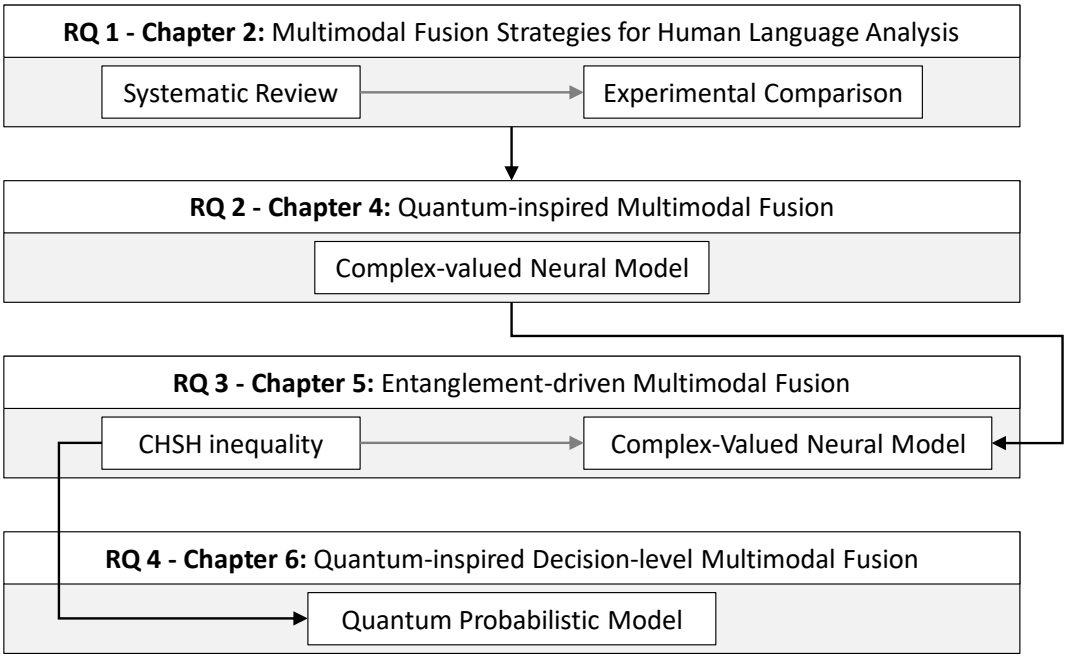


Figure 1.5: Thesis chapters addressing specific research questions.

RQ 1 What are the existing SOTA modality fusion approaches, which components/aspects in these approaches are the most effective to solving the problem, and what are their limitations?

Video sentiment analysis is a rapidly growing area. A recent trend has been geared towards different modality fusion models utilizing various attention, memory and recurrent components [80, 106, 126, 143, 151, 152]. However, there is a lack of systematic investigation on how these different

components contribute to solving the problem as well as their limitations. Chapter 2 answers the first research question by presenting a large-scale and comprehensive empirical comparison of SOTA modality fusion approaches for analyzing human language. The findings provide helpful insights and guidance to the development of more effective quantum probabilistic computational models and identify biases of multimodal datasets.

RQ 2 If and how can we effectively exploit the mathematical formalism of quantum mechanics outside physics to fuse inputs of multimodal features for a video sentiment analysis task?

Existing quantum probabilistic neural models have been investigated for textual representation learning tasks only [78, 139]. On the other hand, current quantum-inspired modality fusion strategies vaguely borrow various detached quantum concepts at different stages, are implemented by real-valued components, and supports only two modalities [163, 164]. They hence neither support trimodals (all three modalities) nor facilitate an end-to-end training. It is an open research question on how the quantum measurement postulate could be deployed outside physics, and especially in the field of affective computing to fuse multimodal features. A significant milestone towards that direction is discussed in Chapter 4 to answer the second research question, by presenting an end-to-end complex-valued neural network, which simulates the quantum measurement postulate, for video sentiment analysis.

RQ 3 How can we encode cross-modal information in the form of non-classical correlations and how such correlations could benefit multimodal decision making?

Recent advances in quantum probabilistic neural models have been shown to achieve a comparable performance to the SOTA, with a better transparency and increased level of interpretability. However, they treat quantum states as either a classical mixture or as a separable tensor product, without triggering their interactions which could make them correlated or non-separable (i.e., entangled in QT term). Such non-separability is a form of non-classical correlation and has been shown in cognitive science as a fundamental feature of human decision making under uncertainty. It complies with the more general quantum probability theory only. This means that the current quantum probabilistic neural models have not fully exploited the expressive power of quantum probabilities. We give a detailed answer to the third research question in Chapter 5, which exposes an extensive investigation on encoding cross-modal information in the form of non-classical correlations. In particular, this chapter reveals the investigation of non-classical correlations in a multimodal Information Retrieval (IR) task, a video sentiment analysis task, and a conversational emotion recognition task. Additionally, various measurements of the degree of non-separability are proposed to optimize post-hoc interpretability in terms of how different modalities interact with each other and contribute to the final decision.

RQ 4 How can we represent users' cognitive states, underlying their multimodal decision perspectives, and how we can learn such representations from multimodal data?

In QT, there is the concept of *incompatibility*, where it is not possible to construct a joint probability distribution for events, and instead we can only assign probabilities to the sequence of outcomes. That is, the outcome of an event is impossible to be in a definite state with respect to two incompatible events because the definite state for one is an indefinite state for another. Incompatibility can also manifest in the individual factors contributing to different dimensions of document relevance in an IR scenario [131] or physiological experiments [41]. Currently, such strategies build up users' cognitive states in a complex-valued Hilbert space, learnt by measuring joint probabilities in user studies. In Chapter 6, we answer the last research question, by proposing a novel approach that constructs the complex Hilbert space representation of the user's cognitive states underlying their multimodal decision perspectives. Moreover, a method for learning such representation from multimodal data has been proposed.

1.5 Main Contributions

The main contributions of this thesis can be summarised as follows:

- We present the first large-scale and comprehensive empirical comparison of eleven SOTA modality fusion approaches in two multimodal human language analysis tasks, with three SOTA benchmarking datasets. A novel methodology is proposed to investigate the effectiveness of SOTA modality fusion strategies for human language analysis and the trade-off between effectiveness and efficiency.

- We introduce the first end-to-end quantum probabilistic fusion neural model, which recognizes the sentiment of a multimodal sentence. In particular, we make the following contributions: 1) we propose a quantum-inspired methodology to model intra-modal interactions between different semantic units, i.e., words, 2) we introduce a strategy to model inter-modal interactions across different modalities deriving inspiration from QT, 3) we designate how to extract the utterance's sentiment in a quantum manner, 4) the quantum view bring a unique superiority in understanding the contributions of single modalities and bi-modals (i.e., any of two modalities) to the predicted sentiment judgment, without requiring an ablation test, and 5) the proposed model achieves a comparable performance to the SOTA.

- We conceive a methodology to formulate the relevance of documents in the form of Bell-type inequalities for a multimodal IR scenario. Bell-type inequalities are a formal method to establish non-classical correlations in data generated by systems. While cognitive scientists have explored ways to use Bell-type inequalities for other human decision-making data [5, 39], it has not been used for multimodal IR data before. This thesis devises a way to formulate relevance judgement data from both image and text unimodal decisions in terms of Bell-type inequalities.

- We propose a transparent and joint quantum probabilistic neural model, which models cross-modal correlations as non-separable, which to our best knowledge, no existing models in the current literature have taken into account. Moreover, we make the following contributions: a) in contrast to previous work, we consider both context modelling, in the form of preceding utterances, and multimodal fusion in the form of modality interactions, into a unified framework, b) the proposed model architecture supports multiparty conversations without requiring artificial expansion, c) the model achieves an improved performance, as compared to various SOTA approaches, for video emotion recognition in conversations, and d) the degree of non-separability of modalities unearths useful and explainable knowledge about the way distinct modalities interact with each other.

- We devise a method for extracting Hilbert Space structure from multimodal data. Hilbert space structure has been used in IR for smaller-scale user study data or as toy examples [17, 43]. However, it has not been exploited for multimodal representation learning tasks. In this dissertation, we introduce a method to represent uni-modal decisions in a Hilbert space with different bases for each modality. Hilbert space is the building block in the quantum framework and lays the foundation for quantum probabilistic models for representation and prediction. It can also be used to test further non-classical phenomena, such as contextuality and interference of individual decisions. The methods developed to model uni-modal decisions into a Hilbert space representing multimodal sentiment judgements is a novel contribution of the thesis.

1.6 Thesis Outline

This dissertation aims at developing transparent and unified quantum probabilistic computational models for fusing different modality streams, borrowing concepts from quantum cognition [25]. All proposed computation models are designed in a way that facilitates accuracy optimization, transparency in the decision-making process, and post-hoc interpretability, giving meaningful explanations to the final decision made.

In **Chapter 2** we present a large-scale and comprehensive empirical comparison of SOTA modality fusion approaches in two video sentiment analysis tasks, with three SOTA benchmarking datasets [49]. A systematic investigation shows how different neural components contribute to solving the problem as well as their limitations. The findings provide us helpful insights and guidance to the development of more effective quantum probabilistic neural models for multimodal fusion.

In **Chapter 3** we present related work which exploits the mathematical formalism of quantum mechanics for representation learning and establish the foundations of QT.

Based on the findings of the Chapter 2 and the limitations of existing quantum-inspired representation learning approaches, discussed in Chapter 3, in **Chapter 4**, we develop an end-to-end complex-valued neural network, simulating the quantum measurement postulate and its procedural

steps for video sentiment analysis [75]. That is, given an utterance constituted of linguistic, visual, and acoustic modalities, the quantum-inspired neural model infers its sentiment, i.e., positive or negative. The model first converts specific-modality features to their quantum analogue, i.e., quantum pure states, and constructs tri-modal representations via the tensor product of specific-modality quantum states. Hence, each word is represented as a unified pure quantum state, which is a unit vector in a complex-valued Hilbert space. Interactions among words in a sentence are modelled into a mixture matrix by calculating the weighted sum of individual words, represented as pure quantum states so that the output to be a valid mixed quantum state. Finally, a set of parameterized vectors applied to the sentence representation by calculating their inner product, as a simulation of the quantum measurement postulate, to identify the discriminating information for sentiment classification.

A step beyond the work in Chapter 4 is the consideration of evolution in quantum systems. To this end, in **Chapter 5** we address a limitation in the literature that quantum probabilistic models treat quantum states as either a classical mixture or as a separable tensor product across modalities, without triggering their interactions in a way that they are correlated or *non-separable*. Such non-separability has been shown in cognitive science as a fundamental feature of human decision making under uncertainty. To this end, we investigate encoding of cross-modal information in the form of non-classical correlations, a.k.a., *entanglement*. In particular, we first present a user study investigating non-classical correlations between image and text for a multimodal IR task, based on the combination of uni-modal decisions [53]. The results provide us with theoretical and empirical insights for the development of a transparent end-to-end probabilistic neural model for video emotion detection in conversations, encoding information in non-classical correlations [51]. The model takes multimodal information, i.e., linguistic, visual, and acoustic, for a sequence of utterances and converts it to quantum states. The concept of *quantum evolution* is first exploited to capture contextual information in the form of preceding utterances. Then the model operates utterance-level pairwise fusion of modalities, i.e., *linguistic-visual*, *linguistic-acoustic*, and *visual-acoustic*, via the tensor product of bi-modals (any of two modalities). The bipartite quantum-states of bi-modals are evolved, so that be entangled, before mapping the complex-valued representation to a real-valued high-level representation via the quantum measurement postulate. Finally, we evaluate the introduced quantum probabilistic model on utterance-level (i.e., without considering preceding utterances) sentiment analysis tasks, modifying the proposed architecture, accordingly [52].

The results of the user study in Chapter 5 show that there is no violation of Bell-type inequalities. Careful thought reveals that one of the reasons for the no violation is because we exploited the classical real-valued representations, but the non-classical properties of QT such as interference and entanglement are rooted on a complex Hilbert Space representation. To this end, in **Chapter 6** we introduce a novel probabilistic model of a highly theoretical and practical significance that constructs the complex-valued Hilbert space representation of user's cognitive states [50], underlying their

multimodal decision perspectives. Afterwards, we propose a method for learning such representation from multimodal data and how they interfere with each other in decision making. The representation is finally used to facilitate a decision fusion approach for utterance-level video sentiment analysis.

Finally, in **Chapter 7** we identify the remaining challenges and discuss the path forward.

1.7 Relationship to Published Work

The chapters in this thesis describe work that has been published or is under review in the following conferences and journals:

Chapter 2

- **Dimitris Gkoumas**, Qiuchi Li, Christina Lioma, Yijun Yu, and Dawei Song. What makes the difference? An empirical comparison of fusion strategies for multimodal language analysis. Information Fusion, 66:184-197, 2021

Chapter 4

- Qiuchi Li, **Dimitris Gkoumas**, Christina Lioma, and Massimo Melucci. Quantum-inspired multimodal fusion for video sentiment analysis. Information Fusion, 65:58–71, 2020.

Chapter 5

- **Dimitris Gkoumas**, Sagar Uprety, and Dawei Song. Investigating non-classical correlations between decision fused multi-modal documents. In International Symposium on Quantum Interaction, pages 163–176. Springer, 2018.
- **Dimitris Gkoumas**, Dawei Song, Qiuchi Li, and Massimo Melucci. An entanglement-driven fusion neural network for video sentiment analysis. In The Web Conference 2021 (Under Review), 2021.
- **Dimitris Gkoumas**, Qiuchi Li, Massimo Melucci, Nie Jian-Yun, and Dawei Song. An entanglement-driven neural network model for contextual and non-separable modality fusion in conversational emotion recognition. Information Fusion (Under Review), 2021.

Chapter 6

- **Dimitris Gkoumas**, Qiuchi Li, Massimo Melucci, and Dawei Song. (2021, February). A quantum cognitively motivated decision fusion for video sentiment analysis. In AAAI 2021.

The publications below describe joint work that is loosely related to this thesis but not described in the thesis:

- Qiuchi Li, **Dimitris Gkoumas**, Alessandro Sordoni, Jian-Yun Nie, and Massimo Melucci. (2021, February). Quantum-inspired neural network for conversational emotion recognition. In AAAI 2021.
- Sagar Uprety, **Dimitris Gkoumas**, and Dawei Song. A survey of quantum theory inspired approaches to information retrieval. ACM Computing Surveys (CSUR), 53(5):1–39, 2020.
- Sagar Uprety, **Dimitris Gkoumas**, and Dawei Song. Investigating bell inequalities for multidimensional relevance judgments in information retrieval. In International Symposium on Quantum Interaction, pages 177–188. Springer, 2018.

Chapter 2

State-of-the-art Multimodal Fusion Approaches

This chapter provides an overview and empirical comparison of the state-of-the-art multimodal fusion approaches for human language analysis, in particular video sentiment analysis. Multimodal video sentiment analysis is a rapidly growing area. A recent trend has been geared towards complex modality fusion models utilizing various attention, memory and recurrent components. However, there lacks a systematic investigation on a) how effective the current machine learning-based multimodal fusion strategies are, b) how efficient the SOTA multimodal fusion strategies are, c) how the effectiveness could affect efficiency and d) which components/aspects are the most effective for affective computing. This chapter fills the gap by first providing a systematic review of multimodal time series for video sentiment analysis and emotion recognition. Then, we experimentally compare the SOTA modality fusion approaches in two video sentiment analysis tasks, with three SOTA benchmarking datasets.

2.1 Related Work

Affective computing. The work in this dissertation targets at computational methodologies for detecting the sentiment or emotion polarity of user-generated videos, a.k.a., video sentiment analysis or emotion recognition, respectively. Both video sentiment analysis and emotion recognition are sub-fields of affective computing [112], which is an emerging interdisciplinary area in multimedia information processing, bringing together AI and cognitive science. In particular, it studies a speaker's sentiment or emotion expressed by verbal (i.e., linguistic) and non-verbal (i.e., visual, acoustic) streams.

Multimodal fusion. One of the major challenges of video affective analysis is to modelling interactions across distinct modalities [116], e.g., linguistic, visual, and acoustic. Early work aggregated decisions of uni-modal classifiers by voting [98], averaging [120], weighted sum [54], classification-based methods [11], e.g., support vector machines, or a trainable model [29, 48, 137]. Other strategies fused features after being extracted [67, 72]. Later work derived inspiration from effective neural strategies in Natural Language Processing (NLP), such as RNN [96], Long Short-term Memory (LSTM) [62], and Convolutional Neural Network (CNN) [69] architectures, to model interactions across different modality features by fusing either input features per timestamp or uni-modal output hidden units [113, 114, 141, 156].

Due to recent advances in deep learning, a recent trend was gearing towards sophisticated RNN backbone architectures. Early advancements in the field utilized tensor-based fusion approaches to compose [151] or factor [15, 85] different modalities at an utterance level [15, 85, 151], word level [79], or in a hierarchical manner [89]. Recently, Mai et al. [91] exploited the tensor-based strategy to fuse segmented uni-modal information for capturing local interactions. Then, the local tensors fed a bidirectional skip connection LSTM to learn global interactions. Other approaches used hybrid memory components, constructed from the hidden units of each modality at the previous timestamp and fed as an additional input of the next timestamp [80, 152]. Inspired by successful trends in NLP, some approaches introduced encoder-decoder structures in sequence-to-sequence learning by translating a target modality to a source modality [38, 90, 106]. Other fusion strategies incorporated reinforcement learning [32], fuzzy logic[30], bilinear pooling [165], deep canonical correlation analysis [124], and hierarchical fusion strategies [46, 57]. Attention mechanisms have also been exploited to align different modalities, resulting in better-performing modality fusion approaches [57, 126, 153, 154]. Currently, attention-based approaches are the most effective fusion approaches for utterance-level human language analysis [49].

Unaligned fusion strategies. One common way of tackling unaligned multimodal sequence is by forced word-aligning before training [107, 126, 113, 152, 157]: manually preprocess the visual and acoustic features by aligning them to the resolution of words. However, it is worth noting that there exist other modality fusion approaches, which manipulate directly unaligned data to elicit long-range contingencies across different modalities (i.e., asynchronous cross-modal interactions) [57, 126, 143]. A fair comparison between word-aligned sequences and unaligned multimodal time series showed a decreased performance for unaligned multimodal streams [126].

Context modelling. Another important aspect of video affective computing is the consideration of context from preceding utterances [116]. Initial work utilized Recurrent Neural Network (RNN) architectures to model interactions between the target utterance and its content [47, 113]. Memory networks were introduced as a more elaborate approach to pay attention to intra-speaker and inter-speaker interactions. In particular, memory cells were deployed to model speaker-specific context [59, 60]. More recent work demonstrated a better level of model transparency and improved performance

utilizing modules to handle intra- and inter-speaker emotional influence [48, 92]. In particular, DialogueRNN [92] built a hierarchical multi-stage RNN with different strategies for updating a speaker and a listener's emotion states. DialogueGCN [48] captured the relations of all utterances in a conversation, based on their relative order and whether they belonged to the same speaker. The relations were reflected in a graph, and a graph neural network was built to update utterance representations. Despite those approaches modelled intra- and inter-speaker interactions effectively, they suppressed cross-modal interactions. In particular, they joined pre-trained utterance-level uni-modal embedding by concatenation or attention mechanism to obtained multimodal embedding and fed it into neural models without eliciting cross-modal interactions. Additionally, current approaches either do not consider any speaker level dependency [47, 113], or support conversations with two speakers only, without can be extended to apply on multiparty conversations [48, 92], or require artificial expansion to be applied on multiparty datasets [92].

Interpretability and information fusion. Most of the current modality fusion strategies are black-box approaches, which come with the price of lacking interpretability. However, recently a few holistic frameworks tried to separate cross-modal interactions, to endow AI models with a better level of interpretability. For instance, the contributions to the prediction from each modality and the interactions between modalities, i.e., bi-modal and tri-modal interactions, have been investigated through an interpretable multimodal fusion framework [127]. Hazarika et al. [61] exploited two subspaces, a joint subspace and a modality-specific subspace, to capture uni-modal and tri-modal interactions. In [154], authors applied seven distinct self-attention mechanisms to a factorized multimodal representation, capturing all possible uni-modal, bi-modal, and tri-modal interactions, simultaneously.

2.2 Methodology for Empirical Comparison of SOTA Models

This section details the methodology we used for our empirical study of the most recent SOTA multimodal language fusion approaches, in the context of video sentiment and emotion analysis tasks. We first formulated the task on which our study was carried out. Sentiment analysis was a binary multimodal classification task inferring either positive or negative emotions. Emotion recognition was a multimodal multilabel classification task inferring one or more emotions, i.e., *neutral*, *happiness*, *sadness*, and *frustration*. However, both tasks aim to capture emotions of video utterance and fall under affective computing field [28].

2.2.1 Task Definition

The goal is to infer the emotion of utterances from video speakers. Each video consists of N sequential utterances $U = (U_1, ..., U_i, ..., U_N)$, where i is the i^{th} utterance. Each utterance U_i is associated with three modalities, namely, linguistic, visual, and acoustic, $U_i = (U_i^l, U_i^v, U_i^a)$, $1 \leq i \leq N$. The

corresponding labels for the N segments are denoted as $y = (y_1, ..., y_i, ..., y_N)$, $y_i \in \mathbb{R}$. We apply word-level alignment, where visual and acoustic features are averaged across the time interval of each spoken word. Then, we zero-pad the utterances to obtain time-series data of the same length. After this step, language, visual, and acoustic features have the same length L. For the linguistic modality the U_i utterance is represented by $U_i^l = (l_i^1, ..., l_i^L)$. Similarly for visual and acoustic modalities, it is represented by $U_i^v = (v_i^1, ..., v_i^L)$ and $U_i^a = (a_i^1, ..., a_i^L)$, respectively.

2.2.2 Datasets and Feature Extraction

We empirically evaluated the SOTA approaches from the last two years on multimodal sentiment analysis tasks by using two SOTA benchmarking datasets, namely CMU Multimodal Opinion-level Sentiment Intensity (CMU-MOSI) [155] and the largest available dataset for multimodal sentiment analysis, CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) [157]. Figure 2.1 illustrates a few samples of monologue video clips. We also evaluated the approaches to the multimodal emotion recognition task using the IEMOCAP dataset [27]. Table 2.1 summarizes the statistics of the datasets in terms of training, validation and test sets. We compared all approaches to word-aligned multimodal language sequences, leaving the very challenging comparison with unaligned language sequences for future work.

Dataset	**Train**	**Validation**	**Test**
CMU-MOSI [155]	1,284	229	686
CMU-MOSEI [157]	16,265	1,869	4,643
IEMOCAP [27]	2,717	798	938

Table 2.1: Training, validation and test data distribution in CMU-MOSI, CMU-MOSEI, and IEMOCAP, respectively.

CMU-MOSI is a relatively balanced (1176 positive and 1023 negative utterances) human multimodal sentiment analysis dataset consisting of 2,199 short monologue video clips (each lasting the duration of a sentence). It has 1,284, 229, and 686 utterances in training, validation, and test sets, respectively. CMU-MOSEI is a larger scale sentiment and emotion analysis dataset made up of 22,777 movie review video clips from more than 1,000 online Youtube speakers. The training, validation, and test sets are comprised of 16,265, 1,869 and 4,643 utterances, respectively. Human annotators labelled each sample with a ratio score from -3 (highly negative) to 3 (highly positive) including zero. Hence, the multimodal sentiment analysis task can be formulated as a regression problem.

For CMU-MOSI and CMU-MOSEI, we used the CMU-Multi-modal Data SDK[1] [157] for feature extraction. Following previous work [85, 126, 143, 151, 152], we converted video transcripts into 300-dimensional pre-trained Glove word embeddings (glove.840B.300d) [104]. Besides, GloVe embedding

[1]https://github.com/A2Zadeh/CMU-MultimodalSDK

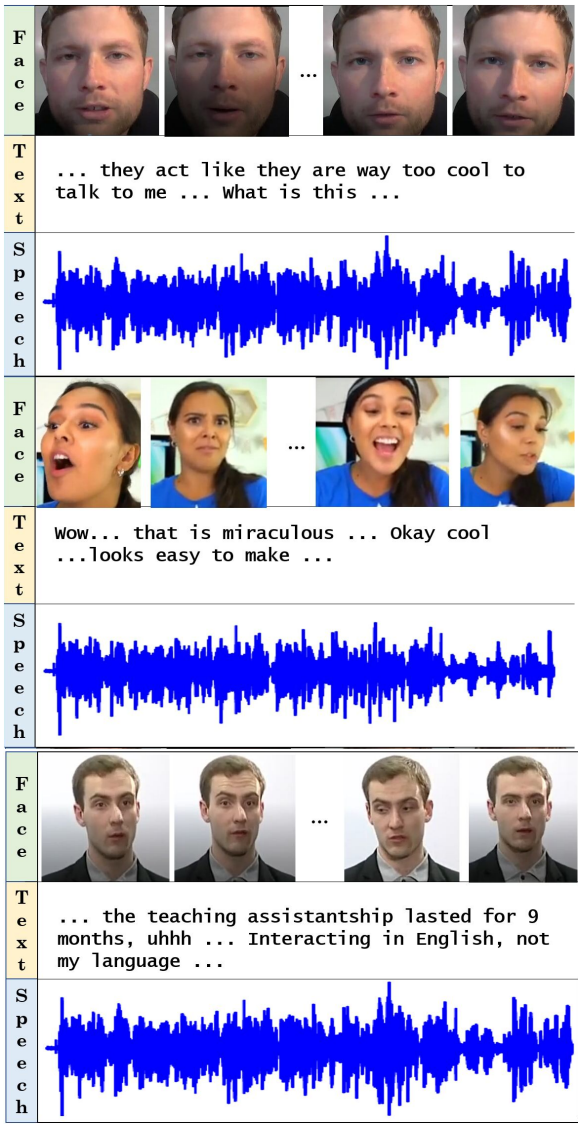


Figure 2.1: Three examples of monologue video clips. The first and third cases corresponds to negative sentiments and the second one to positive sentiment. The task is to predict the sentiment of each utterance for a given set of modalities, i.e., text, visual, and speech, without considering preceding utterances.

is more computationally affordable than other more effective, yet computationally expensive, word embeddings [37, 105]. Facet [2] is used to capture facial muscle movement, including per-frame basic and advanced emotions and facial action units. We used VOCAREP [36] to extract low-level acoustic features (e.g., 12 Mel-frequency cepstral coefficients, pitch tracking and voiced/unvoiced

[2]https://pair-code.github.io/facets

segmenting features, glottal source parameters, peak slope parameters, and maxima dispersion quotients). For CMU-MOSI, we extracted visual and acoustic features at a frequency of $15Hz$ and 12.5 Hz respectively. For CMU-MOSEI, we extracted at a frequency of 15 Hz and $20Hz$. To reach the same time alignment across modalities, we applied a word-level alignment. To align visual and acoustic modalities with words, we used P2FA [150]. Then, to obtain the aligned timesteps, we averaged the visual and audio features within these time ranges. All sequences in the word-aligned case had length 50. For each word the dimension of the feature vector was set to 300 (linguistic), 20 (visual), and 5 (acoustic) for CMU-MOSI, and 300 (linguistic), 35 (visual), and 74 (acoustic) for CMU-MOSEI.

For multimodal emotion recognition, we used IEMOCAP [27]. It consists of 151 videos about dyadic interactions, where professional actors are required to perform scripted scenes that elicit specific emotions. It has 2,717, 798, and 938 utterances in training, validation, and test sets, respectively. Human annotators labelled each sample for four emotions (neutral, happy, sad, angry). The labels for every emotion are binary. That allowed us to reduce the multiclass learning problem to a problem solvable using binary classifiers. Following a one-vs-all strategy, for each emotion, we trained a robust classifier to recognize one emotion from all the others. We followed a similar process to the sentiment analysis datasets to extract features from 3 streams. The linguistic, facial and acoustic embeddings are 300-dimensional, 35-dimensional, and 74-dimensional vectors, respectively. All sequences are word-aligned having length 50.

2.2.3 Evaluation Metrics

To evaluate the effectiveness on CMU-MOSI and CMU-MOSEI tasks, we adopted a series of evaluation performance metrics used in prior work [80, 126, 152, 157]: binary accuracy (i.e., Acc_2 : positive sentiment if $values \geq 0$, and negative sentiment if $values < 0$), 7-class accuracy (i.e., Acc_7 : sentiment score classification in $Z \cap [-3, 3]$), $F1$ score, Mean Absolute Error (MAE) of the score, and the Pearson's correlation ($Corr$) between the model predictions and regression ground truth. For all the metrics, higher values denote better performance, except MAE where lower values denote better performance.

To evaluate the effectiveness on IEMOCAP, in contrast to previous work reporting accuracy [126, 143], we reported recall and F_1 score for individual emotion classes. We empirically found that accuracy was a misleading measurement for evaluating one-vs-all emotion classifiers. That is because there is a class imbalance. For instance, the ratio of utterances labelled as happy versus the other emotion equals 1/6. Indeed, some classifiers showed high accuracy even though they failed to distinguish the emotion of the class from all the others correctly. To evaluate the overall performance of the SOTA models, we also calculated the weighted recall and weighted F_1 score measurements.

We evaluated efficiency by reporting: the number of parameters for each approach, the training time of learning, i.e., speed-up during inference, and the validation set convergence.

2.2.4 Experiments

In this study, we devised three main experiments as follows:

1. **Experiment 1:** We first replicated the SOTA approaches following the same experiment set up, as reported in the original papers. Then, we investigated the performance through a comprehensive critical and experimental analysis.

2. **Experiment 2:** We compared the SOTA approaches in terms of efficiency.

3. **Experiment 3:** We conducted several ablation studies to understand a) the importance of modalities and b) which components contribute most to modelling cross-modal interactions across the three modalities.

2.2.5 SOTA models

We replicated a variety of sequential attention mechanisms, memory, tensor fusion, and translation neural approaches[3] into a unified framework in PyTorch. Most of their authors have made implementations available on Github. We replicated the EF-LSTM, LF-LSTM, RMFN, and MARN models from scratch.

Except for the Multimodal Transformer (MulT) [126], the rest of the modality fusion methods are typically RNN-based deep learning networks. However, we went beyond a typical one-to-one comparison and proposed a taxonomy in terms of model features, namely: recurrent-based, tensor-based, attention mechanism-based, memory-based, and translation-based networks. This taxonomy will enable researchers to understand the SOTA field better and identify directions for future research.

Recurrent cell-based networks

This category includes modality fusion approaches which mainly utilize recurrent cells for each time step. In this case, the cells get stacked one after the other, implementing an efficiently stacked RNN.

- **Early-Fusion LSTM (EF-LSTM)** [62] EF-LSTM concatenates linguistic, visual, and acoustic features at each timestamp, and builds an LSTM to construct sentence-level multimodal representation. The last hidden state is taken and sequentially passed to two fully connected layers to produce the output sentiment.

- **Late-Fusion LSTM (LF-LSTM)** [62]. LF-LSTM builds LSTMs for linguistic, visual, and acoustic inputs separately, and concatenates the last hidden state of the three LSTMs as sentence-level multimodal representation. The concatenated hidden states are taken and sequentially passed to two fully connected layers to produce the output sentiment.

[3]The code for our models and experiments is available on https://github.com/gkoumasd/MSAF

- **Recurrent Multistage Fusion Network (RMFN)** [47] RMFN models cross-modal interactions through a divide-and-conquer approach in several stages. Intramodal dynamics are modelled through modality-specific RNNs. For each timestep, the uni-modal hidden states of RNNs are concatenated. Then, the concatenated representation is processed in multiple stages. For each stage, the most important modalities are highlighted using an attention module and then fused with the previous stage fused representations. In the end, a summary action generates a multimodal joint representation which is fed back into the intramodal RNNs as an additional input for the next timestep.

Tensor-based networks

This group of networks is mainly based on the tensor product of modalities for composing and factorizing information.

- **Tensor Fusion Network (TFN)** [151] TFN explicitly models view-specific and cross-view dynamics by creating a multi-dimensional tensor that captures uni-modal, bi-modal, and tri-modal interactions across linguistic, visual, and acoustic modalities.
- **Low-rank Multimodal Fusion (LMF)** [85]. LMF adopts the same approach as TFN to model the multimodal representation. After that, it applies a tensor decomposition approach by calculating the inner product of the multimodal tensor with a weight tensor. The output is a low-dimension vector, which is used to make predictions.

Attention mechanism-based networks

These approaches mainly exploit various attention mechanism components to fuse modalities.

- **Multi-Attention Recurrent Network (MARN)** [153]. MARM captures cross-modal dynamics at each timestamp. A multi-attention block is built to construct a cross-modal representation, based on hidden states of the previous timestamp, and fed into the inputs of the current timestamp. The cross-modal representation and hidden states of the last timestamp are concatenated to form a multimodal sentence embedding, which is sequentially passed to two fully connected layers to produce the output sentiment.
- **Multimodal Transformer (MulT)** [126] MulT merges multimodal time-series via a feed-forward fusion process from multiple directional pairwise cross-modal transformers. Each cross-modal transformer is a deep stacking of several cross-modal attention blocks. As a final step, it concatenates the outputs from the cross-modal transformers and passes the multimodal representation through a sequence model to make predictions.
- **Multimodal Uni-Utterance - Bi-modal Attention (MMUU-BA)** [47] MMUU-BA encodes linguistic, visual, and acoustic streams through three separate Bi-GRU layers followed by

fully connected dense layers. Then, pairwise attentions are computed across all possible combinations of modalities, i.e., linguistic-visual, linguistic-acoustic, and visual-acoustic. Finally, individual modalities and bi-modal attention pairs are concatenated to create the multimodal representation, used for final classification. MMUU-BA makes predictions by applying a fully connected layer to each timestamp. In our experiments, since we did not consider preceding utterances, we extracted the last hidden state only and fit it to a fully connected layer to make predictions.

- **Recurrent Attended Variation Embedding Network (RAVEN)** [143] RAVEN learns multimodal shifted word representations conditioned on the visual and acoustic modalities. Concretely, visual and acoustic embeddings interact with each word embedding through an attention gated mechanism to yield a nonverbal visual-acoustic vector. The resulting vector is integrated into the original word embedding to model the intensity of the visual-acoustic influence on the original word. By applying the same method for each word in a sentence, the model outputs a multimodal shifted word-level representation. The representation is encoded into an LSTM followed by a fully connected layer to produce an output that fits the task. Yet, in our experiments, we considered the last hidden state to construct nonverbal visual-acoustic embedding since we worked on word-level aligned data.

Memory-based networks

This category extends recurrent neural models with a memory component to model modality interactions.

- **Memory Fusion Network** (MFN) [152] MFN is a memory fusion network that builds a multimodal gated memory component. The memory cell is updated along with the evolution of the hidden states of three uni-modal LSTMs. The last memory cell is concatenated with the last hidden states of uni-modal LSTMs to construct the multimodal sentence representation. Then, the multimodal representation is sequentially passed to two fully connected layers to produce the output sentiment.

Translation-based networks

This category includes neural machine translation approaches for modelling human language by converting a source modality to a target modality.

- **Multimodal Cyclic Translations Network (MCTN)** [106] MCTN is a hierarchical neural machine translation network with a source modality and two target modalities. The first level learns a joint representation by using back translation. Then, the intermediate representation

is translated into the second target modality without back translation. The multimodal representation is fed into RNN for final classification. For our experiments, the source modality is the linguistic modality.

We first fine-tuned all models by performing a fifty-times random grid search on the hyperparameters. We reported the final settings in Appendix A. After the fine-tuning process, we trained all the models again for 50 epochs, five times. We used Adam optimizer with L1 loss as the loss function for CMU-MOSI and CMU-MOSEI since sentiment analysis is formulated as a regression problem. For IEMOCAP, we used cross-entropy loss since emotion recognition is formulated as a multilabel classification problem. We reported the average performance on the test set for all experiments.

2.3 Results

2.3.1 Effectiveness

In Table 2.2, we see that attention mechanism-based approaches, namely, MulT, MMUU-BA, and RAVEN, exhibited the highest binary accuracy (between 78.2% and 78.7%) on CMU-MOSI. MulT reported just 0.1% higher accuracy than RAVEN. Yet, for Acc_7, RAVEN reported an increased performance of 34.6% as compared to 33.8% for MMUU-BA and 33.6% for MulT. TFN attained the highest accuracy of 34.9% for Acc_7. RAVEN and MMUU-BA reported the highest correlation ($Corr$). Despite the low accuracy, MCTN exhibited the lowest mean absolute error. That might imply that MCTN needs more epochs to converge (we found in [106] that MCTN had been trained for 200 epochs). Overall, RAVEN was the most effective approach on CMU-MOSI task. T-tests did not reveal a significant difference in binary accuracy across all approaches.

There was a discrepancy between the empirical results from our experiments and the reported ones in literature. Specifically, we empirically found lower accuracy for all the SOTA approaches, except RAVEN, which attained an increased accuracy of 78.6% compared to 78% in [143]. A possible reason for the discrepancy between literature and our empirical results may be that different versions of the CMU-MOSI dataset had been used in the published works. Those versions consisted of different feature dimensions and sequence lengths. Another possible explanation for this might be the fine-tuning parameters, which were rarely reported in current work, making reproducibility a particularly tricky task. In the literature, MulT was regarded as the SOTA approach among the 11 investigated approaches, reporting an increased binary accuracy of 83.0% as compared to 78.7% in our experiments on CMU-MOSI. Note that for MulT we used the same datasets, implementation, and configuration settings as described in [126].

In Table 2.3, we present the results for multimodal sentiment analysis on CMU-MOSEI. All approaches attained an improved performance compared to that of the CMU-MOSI dataset. We suspect this is because CMU-MOSEI is a much larger dataset. MMUU-BA attained an increased

Model	Acc_7	Acc_2	$F1$	MAE	$Corr$
LSTM					
EF-LSTM [62]	32.7	75.8	75.6	1.000	0.630
LF-LSTM [62]	32.7	76.2	76.2	0.987	0.624
RMFN [47]	32.3	76.8	76.4	0.980	0.626
Tensor					
TFN [151]	**34.9**	75.6	75.5	1.009	0.605
LMF [85]	30.5	75.3	75.2	1.018	0.605
Attention					
MARN [153]	31.8	76.4	76.2	0.984	0.625
MulT [126]	33.6	**78.7**	78.4	0.964	0.662
MMUU-BA [47]	33.8	**78.2**	78.1	0.947	**0.675**
RAVEN [143]	**34.6**	**78.6**	78.6	0.948	**0.674**
Memory					
MFN [152]	31.9	76.2	75.8	0.988	0.622
Translation					
MCTN [106]	32.3	76.2	76.2	**0.903**	0.630

Table 2.2: Comparative analysis across the SOTA approaches on CMU-MOSI. Best results are highlighted in bold.

Approach	Acc_7	Acc_2	$F1$	MAE	$Corr$
LSTM					
EF-LSTM [62]	45.7	78.2	77.1	0.687	0.573
LF-LSTM [62]	47.1	79.2	78.5	0.655	0.614
Tensor					
TFN [151]	47.3	79.3	78.2	0.657	0.618
LMF [85]	47.6	78.2	77.6	0.660	0.623
Attention					
MARN [153]	47.7	79.3	77.8	0.646	0.629
MulT [126]	46.6	**80.2**	79.8	0.657	0.661
MMUU-BA [47]	**48.4**	**80.7**	80.2	**0.627**	**0.672**
RAVEN [143]	47.8	**80.2**	79.8	0.636	0.654
Memory					
MFN [152]	47.4	79.9	79.1	0.646	0.626

Table 2.3: Comparative analysis across the SOTA approaches on CMU-MOSEI. Best results are highlighted in bold.

binary accuracy of 80.7% compared to 80.2% for RAVEN and MulT. MMUU-BA also reports the highest accuracy for Acc_7 and the highest correlation ($Corr$ in Table 2.3) compared to all other approaches. In general, we found that attention mechanism-based fusion strategies, namely, MMUU-BA, MulT, and RAVEN, significantly outperformed the other approaches. Yet, there is no significant difference across MMUU-BA, MulT, and RAVEN in terms of binary performance.

CMU-MOSEI is a recently published dataset. We can only compare the empirical results from our experiments to the reported ones in literature for RAVEN, MulT and MMUU-BA. In literature, MulT reported the best binary performance, attaining an increased binary accuracy of 82.5% compared to 80.2% in our experiments even though we used the same experimental settings as in [126]. In contrast, MMUU-BA reported an increased binary accuracy of 80.7% compared to 79.8% in literature. In [143], authors did not conduct experiments on CMU-MOSEI. Yet, in [126], for RAVEN, authors reported a decreased accuracy of 79.1% compared to our 80.2% (see Table 2.3). We could not run experiments for RMFN and MCTN on CMU-MOSEI. RMFN was computationally too expensive, and MCTN could not support CMU-MOSEI due to its computational complexity and required memory resources on the large-scale CMU-MOSEI task.

Following previous work [84], the binary performance across different modality fusion approaches was compared for the CMU-MOSI and CMU-MOSEI tasks, as shown in Figure 2.2. Each line style corresponds to the taxonomy of the SOTA approaches. According to Figure 2.2, all approaches improved on the CMU-MOSEI task. Besides, MulT and RAVEN yielded similar performance for both CMU-MOSI and CMU-MOSEI tasks. That is, they showed similar learning behaviour. However, MMUU-BA showed a positive trend with a sharper rise in performance for the CMU-MOSEI task than the MulT and RAVEN approaches.

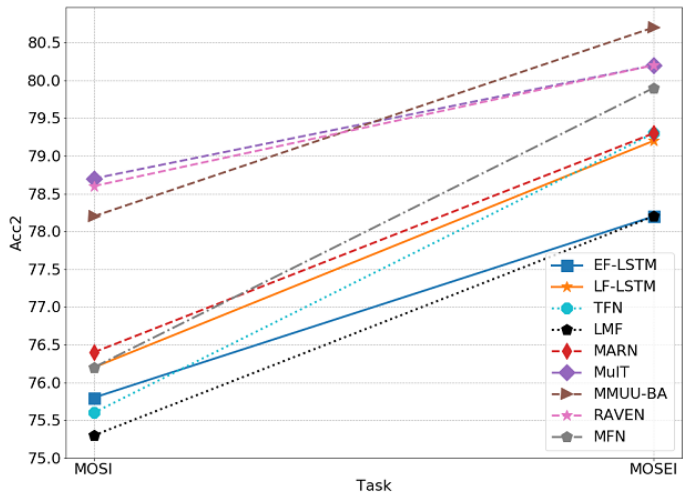


Figure 2.2: Accuracy comparison across different modality fusion approaches for CMU-MOSI and CMU-MOSEI tasks.

We present the results for the emotion recognition task in Table 2.4. In contrast to sentiment analysis tasks, which calculated accuracy, we calculated the class-wise recall to find out how many emotions were detected correctly out of the total number of emotions for each emotion class. We also calculated the weighed recall for each modality fusion method. The results showed that the happy emotion class was the most challenging for all approaches, while the angry class was the most straightforward. Attention mechanism approaches, e.g., MulT and MMUU-BA, were the most effective for the emotion recognition task. In particular, MMUU-BA achieved the highest recall for happy and sad classes, while MulT recalled the most neutral utterances (see Table 2.4). However, EF-LSTM had the highest sensitivity for the angry class. Overall, MulT was the most effective approach for the emotion recognition task, yielding an increased weighted recall of 60.2% as compared to 58.7% of the next best approach, i.e., MMUU-BA. We could not directly compare our results with those in literature since binary accuracy was used as a prime performance measurement. However, in [126], MulT is also the SOTA for the IEMOCAP task.

	Neutral		Happy		Sad		Angry		Weighted	
Approach	*Recall*	*F*1	*Recall*	*F*1	*Recall*	*F*1	*Recall*	*F*1	*Recall*	*F*1
LSTM										
EF-LSTM [62]	57.3	61.2	20.7	30.8	57.7	**62.0**	**80.7**	71.7	57.8	59.5
LF-LSTM [62]	58.5	60.0	31.7	40.0	53.7	56.0	66.1	69.6	55.5	58.6
RMFN [47]	56.9	60.3	17.3	25.6	55.4	57.3	65.5	70.8	53.2	57.2
Tensor										
TFN [151]	60.0	61.9	19.3	28.0	53.4	57.3	76.4	**72.9**	56.7	58.7
LMF [85]	46.6	54.7	34.5	40.6	49.8	54.3	80.1	**72.9**	53.6	57.0
Attention										
MARN [153]	55.1	59.6	27.1	35.1	57.2	57.4	70.4	71.2	55.2	58.4
MulT [126]	**64.9**	**64.2**	19.9	29.6	56.8	58.5	79.3	70.9	**60.2**	59.7
MMUU-BA [47]	57.0	60.0	**35.6**	41.8	**58.2**	61.2	75.5	71.9	58.7	**60.5**
RAVEN [143]	33.6	42.6	0.7	1.4	14.5	23.2	21.4	32.7	22.0	30.3
Memory										
MFN [152]	49.4	55.6	35.1	**42.1**	56.2	55.5	64.5	67.3	52.4	56.5

Table 2.4: Comparative analysis across the SOTA approaches on IEMOCAP dataset. Best results are highlighted in bold.

Overall, we see that all approaches attained a lower binary performance compared to the reported ones in literature, except RAVEN, which achieved higher performance on both CMU-MOSEI and CMU-MOSI, and MMUU-BA, which achieved a higher accuracy on CMU-MOSEI. RAVEN was the most effective model for the CMU-MOSI task, MMUU-BA for CMU-MOSEI, and MulT for IEMOCAP. That is, attention mechanism-based approaches were the most effective for human multimodal affection recognition tasks. MulT was a robust competitive model, but, in contrast to the literature, we found that it did not attain the highest performances on sentiment analysis tasks. Nevertheless, without considering efficiency, we noticed that MulT, MMUU-BA, and RAVEN were the most appropriate models for sentiment analysis, while MMUU-BA and MulT were the most

appropriate ones for emotion recognition. While RAVEN showed outstanding performance for the sentiment analysis tasks, it yielded the lowest performance for the emotion recognition task.

2.3.2 Error Analysis

We conducted an error analysis on the above experiments. Figure 2.3 shows the percent error[4] per sentiment class on CMU-MOSI. Each line style corresponds to the taxonomy of the SOTA approaches. Although CMU-MOSI is a relatively balanced dataset, consisting of 1176 positive and 1023 negative utterances, all fusion modality approaches yielded a higher percent error for the positive sentiment class compared to the negative sentiment class (see Figure 2.3). In particular, most approaches showed a percent error that was twice as high for the positive sentiment class compared to the negative sentiment class. We also noticed that attention mechanism-based approaches, e.g., MMUU-BA, MulT, and RAVEN, achieved the lowest percent error for the positive sentiment class. However, tensor-based modality fusion approaches, e.g., TFN and LMF, were more effective in terms of performance for the negative sentiment class. It is worth noting that RAVEN, achieving the lowest percent error for the positive class, yielded the highest percent error for the negative class.

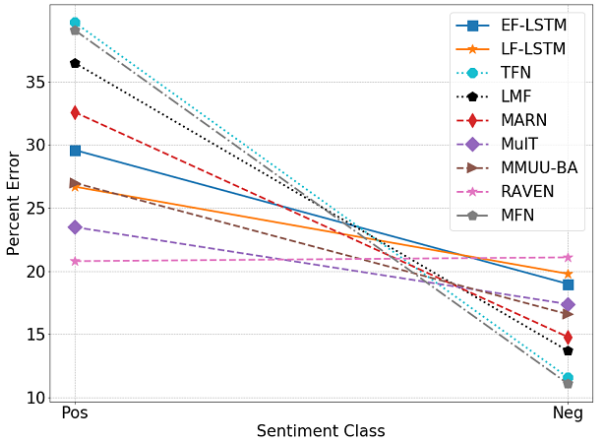


Figure 2.3: Percent error per sentiment class on CMU-MOSI.

Figure 2.4 depicts the percent error per sentiment class on CMU-MOSEI. In contrast to CMU-MOSI, all approaches achieved a low percent error for the positive sentiment class, whereas they struggled with negative utterances. We suspect this was because CMU-MOSEI is an unbalanced dataset. That is, it consists of 11544 positive and 4721 negative utterances. The results showed that once we collected enough data, there was no significant difference among different fusion modality approaches in terms of performance (see the positive class in Figure 2.4).

[4]We define percent error within a class as the difference between the estimated number and the actual number when compared to the actual number expressed as a percentage.

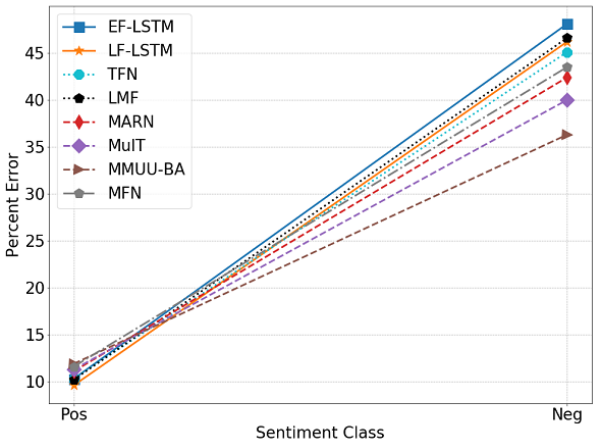


Figure 2.4: Percent error per sentiment class on CMU-MOSEI.

Figure 2.5 shows the percent error for each emotion on IEMOCAP. The results showed that the percent error was high, i.e., greater than 64%, for the happy emotion class. We suppose that this was due to the small number of samples. Specifically, the happy emotion class has only 135 samples compared to 383, 193, and 227 in the neutral, sad, and angry emotion classes, respectively, in the test set. That implies that the performance for each emotion class was analogous to the number of samples for each class. However, some approaches, such as MMUU-BA and MulT, were more effective than others, such as RAVEN and MFN. That is, there was considerable variance in percent error across different modality fusion approaches.

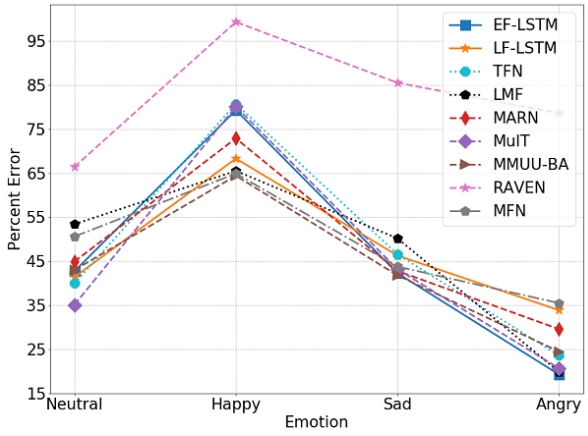


Figure 2.5: Percentage error per emotion class on IEMOCAP.

We then carried out the following analysis on test outputs of CMU-MOSI. We grouped the outputs of all the samples in the test dataset. The first group (i.e., easy) consisted of 49 cases,

where all methods predicted correctly; the second group (i.e., medium) consisted of 21 cases, where half the methods predicted correctly; the third (i.e., hard) consisted of 18 cases, where 2 out of 11 methods predicted correctly; and the fourth (i.e., very hard) consisted of 15 cases, where all methods predicted incorrectly. We included four linguistic-visual samples for each group in Table 2.5.

Out of 686 utterances, 49 of them, that were 7.1%, were predicted correctly by all approaches. These were usually sentences consisting of highly sentimental words such as "horrible", "love" (see Table 2.5, Easy category). Only 21 utterances, 3.1%, were predicted correctly by half of the approaches. All those utterances were either neutral or positive. For example, one possible reason that approaches failed to make a correct prediction for utterances such as "*But it does have some adult humour*" and "*It actually surprised me*" (see Table 2.5, Medium category) was due to missing context and the confused visual content. Eighteen utterances, i.e., 2.6%, could not be correctly predicted by 9 out of 11 approaches, even though utterances included highly sentimental words like "pretty girl", "laughing", but neutral or negative facial expressions (see Table 2.5, Hard category). Finally, no approaches could predict 15 utterances, that was 2.2%. Utterances like "*Everything that happened in Shrek 1,2, and 3 are wiped away*" and "*A lot of people don't like Scream 2*" (see Table 2.5, Very Hard category) were dominated by highly negative words, but the overall sentiment was positive. It is worth mentioning that all the error cases of the medium, hard, and very hard groups were positive sentiment utterances. To our knowledge, this is a novel finding.

2.3.3 Efficiency

In experiment 2, we reported the model sizes (i.e., parameters), the training time of learning, and the validation set convergence. We illustrate the validation set convergence across all competitive approaches on CMU-MOSI, CMU-MOSEI and IEMOCAP in Figure 2.6, Figure 2.7, and Figure 2.8, respectively. We noticed that all approaches converged in just a few epochs for all tasks, i.e., CMU-MOSI, CMU-MOSEI, and IEMOCAP tasks. Overall, we observed that the validation set convergence exhibited different carve trends across different fusion approaches and tasks. At first, all approaches manifested a downtrend. That implies that the learning algorithms sought to minimize the loss function, called optimization. After the optimization process, there were some approaches that the downtrend shifted to an uptrend with a sharp rise (e.g., observe LMF and EF-LSTM convergence in Figure 2.7, or MFN and LF-LSTM in Figure 2.8). We attribute such a sharp rise to overfitting. Indeed, some approaches were more prone to overfitting than others. Other strategies exhibited a horizontal trend (e.g., the majority of models in Figure 2.6, or MULT and MFN in 2.7) after the optimization process. That means that the optimization algorithm was stuck in a local optimal - a good enough set of weights - or a global optimal - the best set of weights. However, for CMU-MOSI task, the horizontal trends were smooth while for CMU-MOSEI task, they usually manifested a slight negative or positive slope. We speculate that this was due to the high learning rate on CMU-MOSEI.

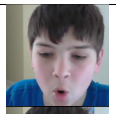
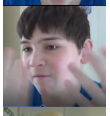
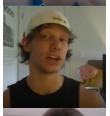
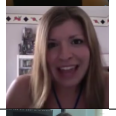
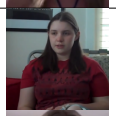
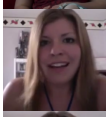
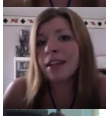
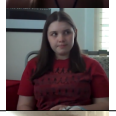
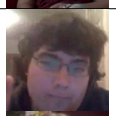
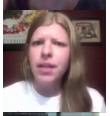
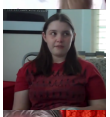
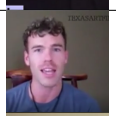
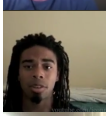
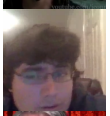
Category	Linguistic	Visual	Sentiment
Easy (100%)	This movie was horrible.		Negative
	I had no idea why I even saw this movie.		Negative
	This movie seemed um a little long.		Negative
	You will really love this movie if you are 8.		Positive
Medium (50%)	But it does have some adult humour.		Positive
	He is pretty average guy.		Positive
	The two women in this movie are particularly good looking.		Positive
	It actually surprised me.		Positive
Hard (20%)	They are back to you having two killers thankfully.		Positive
	She is really pretty girl.		Positive
	It had me laughing out loud.		Positive
	Not bad idea for a sequel.		Positive
Very Hard (0%)	Who I don't usually like		Positive
	I did like Transformers 2 even though a lot of people didn't like that.		Positive
	A lot of people don't like Scream 2.		Positive
	Everything that happened in Shrek 1,2 and 3 are wiped away.		Positive

Table 2.5: Error cases across all approaches on CMU-MOSI task. We illustrate the linguistic and visual parts, which humans can easily understand.

For CMU-MOSI, we empirically found that MMUU-BA converged faster to better results at training compared to other approaches (see Figure 2.6). RAVEN showed a more stabilized mean absolute error (MAE) at training compared to MulT, but it was still higher compared to MMUU-BA. In general, all approaches converged quite fast, up to 10 epochs. We assume that this was due to the small data size. We observed that MCTN needs much more than 50 epochs to converge.

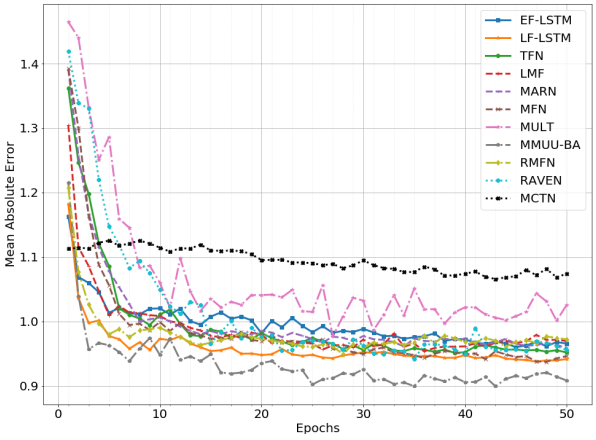


Figure 2.6: Validation set convergence across the SOTA approaches on the CMU-MOSI task

For CMU-MOSEI, we observed that EF-LSTM, LF-LSTM, TFN, LMF, and MARN increased the MAE after 5 epochs (see Figure 2.7). A possible explanation for this might be overfitting since CMU-MOSEI was a large dataset. MulT and RAVEN showed a pretty destabilized MAE at training. Despite RAVEN being among the most robust approaches on CMU-MOSEI in terms of binary accuracy, it achieved the highest MAE among all approaches (see Figure 2.7). Finally, we empirically found that MMUU-BE converged faster to better results, attaining the lowest MAE.

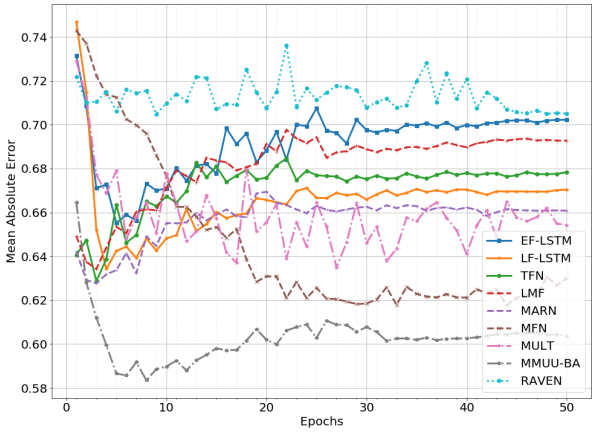


Figure 2.7: Validation set convergence across the SOTA approaches on the CMU-MOSEI task

For IEMOCAP, most of the approaches increased the cross-entropy loss after the 5^{th} epoch (see Figure 2.8). Only RAVEN and MulT attained a low and stabilized cross-entropy loss. Specifically, MulT, reporting the best recall performance for the "neutral" class, attained the lowest cross-entropy loss. EF-LSTM, achieving an improved performance as compared to other sophisticated competitive approaches, showed a fair and stabilized loss at training until 25^{th} epoch.

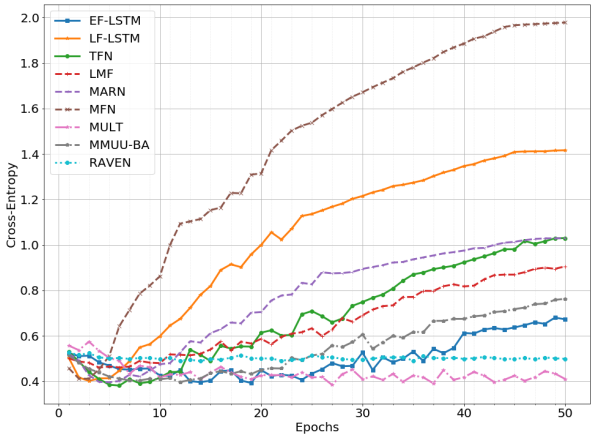


Figure 2.8: Validation set convergence across the SOTA approaches on the IEMOCAP task

We investigated the complexity of the models by presenting the number of parameters and training times in minutes for CMU-MOSI, CMU-MOSEI, and IEMOCAP in Table 2.6. We observed that approaches integrating LSTMCell components, such as LF-LSTM, MARN, and RMFN, were not able to speed up. PyTorch could not maintain the same speed for LSTMCell, which is a variant of LSTM. Despite the low performances, tensor-based approaches attained significant speedup during inference. For CMU-MOSI, MMUU-BA was faster than RAVEN, even though the latter had fewer parameters. We attribute this slowdown to the LSTMCell component of RAVEN. MulT, being a more complicated model, required more time (i.e., 17.6 minutes) than MMUU-BA and RAVEN (i.e., 0.64 and 3.71 minutes, respectively). We observed similar behaviour for CMU-MOSEI. Even though CMU-MOSEI was a relatively large dataset compared to CMU-MOSI, some models had fewer parameters on CMU-MOSEI compared to CMU-MOSI. This might be because different configuration settings were set up after the fine-tuning process. For IEMOCAP, EF-LSTM was not only an effective but also an efficient approach, attaining a more significant speedup (26 times) than its counterpart (i.e., MulT) in terms of performance.

2.3.4 Ablation Studies

To address the third research question, we designed various ablation studies to analyse a) the importance of modalities and b) essential components for learning cross-modal interactions. We conducted

	MOSI		MOSEI		IEMOCAP	
Approach	Mins.	Params.	Mins.	Params.	Mins.	Params.
LSTM						
EF-LSTM [62]	0.43	177,329	6.59	217,457	1.40	206,152
LF-LSTM [62]	3.14	1,155,109	54.47	5,111,485	3.59	946,756
RMFN [47]	57.42	1,950,805	-	-	20.85	1,732,884
Tensor						
TFN [151]	0.51	14,707,911	1.87	6,804,859	0.53	23,198,398
LMF [85]	0.43	1,144,493	2.00	5,079,473	1.12	962,116
Attention						
MARN [153]	69.5	1,350,389	268.20	5,442,313	4.6	1,362,116
MulT [126]	17.6	1,071,211	31.20	874,651	36.89	1,074,998
MMUU-BA [47]	0.64	2,424,965	7.07	2,576,165	0.79	2,605,484
RAVEN [143]	3.71	171,433	23.87	159,213	3.00	173,680
Memory						
MFN [152]	1.88	1,513,221	18.56	415,521	5.13	1,325,508
Translation						
MCTN [106]	15.64	147,100	-	-	-	-

Table 2.6: Comparative analysis across the SOTA approaches on IEMOCAP dataset.

all ablation studies on CMU-MOSI.

Importance of Modalities

To understand the importance of modalities in multimodal tasks, we conducted ablation studies on TFN, which inherently models uni-modal, bi-modal, and tri-modal interactions, and MulT, which attains high accuracy on both sentiment analysis and emotion recognition tasks. For TFN, we tested the TFN approach with uni-modal, bi-modal, and tri-modal tensors. Table 2.7 shows the results of the ablation studies. We observed that language was the most informative modality as it was a pivot for visual and acoustic modalities. The uni-modal visual and acoustic subnetworks and the bi-modal visual-acoustic subnetwork attained fairly low accuracy compared to those integrating the linguistic modality. Specifically, combining language with visual or acoustic modalities was generally better than combining the visual and acoustic modalities. In contrast to [151], we found that the language-based subnetwork performed similarly to the tri-modal tensor network in terms of the binary accuracy. That is, our experiments showed that the tensor-based fusion was not an effective approach for modelling cross-modal interaction across three modalities.

For MulT, we first considered the performances for linguistic, visual, and acoustic only transformers. We found a binary accuracy of 79.5% for the language transformer compared to 77.4% in literature [126]. The language transformer significantly outperformed the visual- and acoustic-only transformers (see Table 2.8).

We also studied the importance of individual cross-modal transformers according to the target

Variant	Acc_7	Acc_2	$F1$	MAE	$Corr$
TFN_l	31.3	**75.7**	75.6	1.017	**0.756**
TFN_v	17.3	53.2	50.5	1.465	0.125
TFN_a	15.2	56.6	54.4	1.425	0.181
$\text{TFN}_{l,v}$	30.3	75.1	75.0	1.013	0.610
$\text{TFN}_{l,a}$	31.1	75.9	75.9	**1.012**	0.624
$\text{TFN}_{v,a}$	15.4	56.9	55.5	1.414	0.178
$\text{TFN}_{w/oc}$	35.7	75.1	74.9	1.024	0.605
$\text{TFN}_{l,v,a}$ [151]	**34.9**	75.6	75.5	1.009	0.605

Table 2.7: Comparison of TFN with with other variants of it on CMU-MOSI.

Variant	Acc_7	Acc_2	$F1$	MAE	$Corr$
MulT_l	**34.3**	**79.5**	79.2	**0.939**	0.662
MulT_v	20.9	59.7	58.3	1.401	0.154
MulT_a	18.75	60.5	60.1	1.348	0.211
$\text{MulT}_{v,a\rightarrow l}$	31.3	76.7	76.5	1.037	0.604
$\text{MulT}_{l,a\rightarrow v}$	32.6	78.9	78.7	0.993	**0.787**
$\text{MulT}_{l,v\rightarrow a}$	33.6	**79.6**	79.4	0.996	0.663
MulT_{H_5}	31.9	79.0	78.8	1.014	0.662
$\text{MulT}_{H_{10}}$	33.5	79.0	79.0	0.995	0.667
MulT [126]	33.6	78.7	78.4	0.964	0.662

Table 2.8: Comparison of MulT with other variants of it on CMU-MOSI.

modality (i.e., $L, V \rightarrow A$, $V, A \rightarrow L$, and $L, A \rightarrow V$). Among the three cross-modal transformers, the one where acoustic was the target modality worked best. This result is consistent with [106] but in contrast with [126], which reports that presenting language as a target modality leads to the best performance. The experiments showed that there was no need to consider multiple directional pairwise cross-modal transformers. Specifically, when we considered acoustic as a target modality yielded an increased accuracy of 79.6% compared to 78.7% for MulT. However, there was no statistical difference in performance among the three cross-modal transformers and the multiple directional pairwise cross-modal transformer (i.e., MulT).

Important Modules for Cross-modal Interactions

To understand the influence of individual components for modelling cross-modal interactions, we performed comprehensive ablation analysis on the SOTA approaches on CMU-MOSI. First, we studied the importance of extra dimensions with value 1 of $\text{TFN}_{l,v,a}$ [151], which models uni-modal and bi-modal dynamics, besides tri-modal ones. We found that the TFN version without constant ($\text{TFN}_{w/oc}$ in Table 2.7) reported a decreased accuracy of 75.1% compared to 75.6% for TFN. However, for Acc_7, the model improved from 34.9% to 35.7% when comparing $\text{TFN}_{l,v,a}$ to $\text{TFN}_{w/oc}$.

For MulT, we considered the number of heads in the cross-modal attention module. We experimented with 5 and 10 heads (MulT_{H_5} and $\text{MulT}_{H_{10}}$ in Table 2.8, respectively). We did not observe any difference in terms of binary accuracy. However, for Acc_7, the increased number of heads yielded an increased performance of 33.5% compared to 31.9% (see Table 2.8).

In [153], authors claim that for each timestamp, there might exist multiple cross-modal interactions. We experimented with three variants of MARN to investigate the number of attentions needed to extract all cross-modal dynamics. Specifically, we tried one, five, and ten attentions. In contrast to [153], our experiments showed that the MARN with only one attention slightly outperformed the models with multiple attentions in terms of binary accuracy (see Table 2.9). Nevertheless, the MARN with five attentions outperformed the other two variants, for Acc_7. We also removed the multi-attention block (MAB) from MARN. Specifically, we replaced the MAB with a fully connected layer and removed the softmax function. We observed that there was no effect on binary accuracy (see Table 2.9) while for Acc_7, the difference was marginal.

Variant	Acc_7	Acc_2	$F1$	MAE	$Corr$
$\text{MARN}_{K=1}$	30.9	**76.9**	76.7	0.983	**0.629**
$\text{MARN}_{K=5}$	31.5	76.1	76.0	1.001	0.616
$\text{MARN}_{K=10}$	30.9	76.4	76.2	1.012	0.621
$\text{MARN}_{w/oMAB}$	**32.4**	76.4	76.2	**0.979**	0.622
MARN [153]	31.8	76.4	76.2	0.984	0.625

Table 2.9: Comparison of MARN with other variants of it on CMU-MOSI.

For MMUU-BA, we analyzed the attention module to understand its learning behaviour. We experimented with two other variants of MMUU-BA (see Table 2.10). The architecture of these variants differed concerning the attention computation module. Particularly, in MMUU-UA, we computed one-directional attention, e.g., from linguistic to visual modality only. In MMUU-SA, we only computed self-attention within modalities. We found that one-directional attention resulted in an increased binary accuracy of 78.8% compared to 78.2% from the proposed framework. Both MMUU-UA and MMUU-BA attained the same performance, for Acc_7 (see Table 2.10). For the self-attention approach, we found that it was less effective than the one-directional cross-modal attention but more effective than the bi-directional cross-modal attention, in terms of the binary performance.

Variant	Acc_7	Acc_2	$F1$	MAE	$Corr$
MMUU-UA	**33.8**	**78.8**	78.6	**0.925**	0.680
MMUU-SA	32.0	78.6	78.5	0.950	**0.688**
MMUU-BA [47]	**33.8**	78.2	78.1	0.947	0.675

Table 2.10: Comparison of MMUU with other variants of it on CMU-MOSI.

For MFN, first, we investigated if cross-modal interactions could happen over multiple time

instances. Specifically, we experimented with a variant of MFN by shrinking the context from time t and $t-1$ to only the current timestamp t in the memory component. We found that $\text{MFN}_{w/o\Delta}$ (see Table 2.11) significantly underperformed the MFN approach. That implies that we should not model cross-modal interactions on aligned time steps, but consider long-range cross-modal contingencies across a multimodal sequence. Second, we evaluated the importance of spatial-temporal cross-modal interactions over time by removing all memory components. The results showed the effectiveness of memory components on the proposed approach. Both outcomes agree with the reported experiments in [152].

Variant	Acc_7	Acc_2	$F1$	MAE	$Corr$
$\text{MFN}_{w/o\Delta}$	31.5	73.8	73.8	1.042	0.584
$\text{MFN}_{w/oMemory}$	31.6	75.0	74.8	1.011	0.598
MFN [152]	**31.9**	**76.2**	75.8	**0.988**	**0.662**

Table 2.11: Comparison of MFN with other variants of it on CMU-MOSI.

For RAVEN, we have already removed the Nonverbal Subnetworks [143] as mentioned in Section 2.2.5. This modification resulted in an increased binary accuracy of 78.6% compared to 78.0% in [143] on CMU-MOSI. We also investigated the temporal interactions between the nonverbal "sub-word" units with language utterances. Specifically, we removed the shift component, which learns dynamically to shift the text representation by integrating the nonverbal vector. Visual and acoustic representations were concatenated with the word embeddings before being fed to downstream networks. We found that integrating the nonverbal context with words was beneficial for understanding human language (see Table 2.12). Specifically, RAVEN showed a significantly increased binary performance of 78.6% compared to 75.6% for $\text{RAVEN}_{w/oShift}$.

Variant	Acc_7	Acc_2	$F1$	MAE	$Corr$
$\text{RAVEN}_{w/oShift}$	31.8	75.6	75.5	1.016	0.615
RAVEN [143]	**34.6**	**78.6**	**78.6**	**0.948**	**0.674**

Table 2.12: Comparison of RAVEN with other variants of it on CMU-MOSI.

For RMFN, we decomposed the fusion problem into multiple stages; we experimented with the number of stages needed for modelling cross-modal dynamics. Specifically, we experimented with one, three, and six stages. Our experiments showed that RMFN attained a similar performance, whether we applied one or six stages to fuse information (see Table 2.13).

Overall, we found that linguistic modality was a pivot for visual and acoustic modalities. This basic finding is consistent with literature. However, the results from ablation studies did not always follow findings reported in literature. In particular, we found that:

- fusing multimodal information into multiple levels (e.g., MulT, MARN, and RMFN) does not

Variant	Acc_7	Acc_2	$F1$	MAE	$Corr$
$\text{RMFN}_{s=1}$	32.9	75.3	75.2	**0.982**	0.616
$\text{RMFN}_{s=3}$	32.5	75.5	75.3	0.991	**0.623**
$\text{RMFN}_{s=6}$	**33.1**	**75.6**	75.5	0.991	0.613
RMFN[47]	31.7	75.2	75.1	1.005	0.612

Table 2.13: Comparison of RMFN with other variants of it on CMU-MOSI.

necessarily result in better binary performance. In some cases, fusing information into multiple levels might achieve slightly better fine-grained accuracy, that is, Acc_7;

- tensor-based approaches underperform the linguistic modality;
- integrating the temporal (e.g., MFN) or modality (e.g., RAVEN) context over the multimodal fusion process results in a significantly better performance.

2.4 Discussion on Key Findings

In this study, we replicated the most recent SOTA models for multimodal language analysis. We evaluated their effectiveness through comprehensive comparative studies, error analyses and series of ablation studies. The efficiency of the models was also compared in terms of three evaluation metrics, namely, parameters, training time, and validation set convergence. The results associated with ablation studies helped us determine which components and methodologies contribute most to solving the problem of affective computing.

In terms of effectiveness, the experiments showed that approaches exploiting attention mechanism components improved the model performance for both sentiment analysis and emotion recondition tasks. We speculate that this was because the attention mechanism acted as an implicit multimodal alignment component. Memory networks reached a similar performance as well. On the other hand, despite tensor-based approaches getting a lower present error for the negative sentiment class on CMU-MOSI, in general, they did not attain high performance. Similarly, recurrent cell-based approaches did not achieve a high performance either. Overall, most of the SOTA approaches attained lower performances in the range of 2% to 4.5% compared to the reported one in the literature. We mainly attribute such discrepancies to the fine-tuning process. The different versions of the CMU-MOSEI and CMU-MOSI datasets used in published works could be another reason for most of those cases.

From an efficiency viewpoint, attention mechanism-based approaches were usually more complex and required more training time than the rest of the modality fusion approaches. To alleviate that issue, we could consider less fine-grained cross-modal interactions. Indeed, our ablation studies showed that adding more levels of interactions across modalities resulted in a decreased performance.

Recurrent cell-based approaches were extremely computationally expensive. On the other hand, memory and tensor networks were more efficient.

Table 2.14 summarizes the key findings on how different components contributed to solving the problem of affective video content analysis. Overall, the results demonstrated that attention mechanism were the most effective approaches despite being computationally expensive. During the training process, they manifested a stabilized and fast convergence, and they coped with both skewed and balanced datasets. However, autoencoder approaches were more suitable for missing or noisy data. The ablation studies showed that cross-modal interactions are not aligned on corresponding time steps but spread across a multimodal sequence. Finally, video sentiment analysis could benefit from the integration of context. However, all approaches struggled with positive sentiment utterances.

Component	**Model**	**Contribution**
Basic Recurrent Structures	EF-LSTM [62], LF-LSTM [62]	1) Computationally cheap. 2) Outperform a few SOTA approaches.
Tensor Operator	TFN [151], LMF [85]	1) Low error for the negative class on MOSI. 2) Computationally cheap.
Attention Mechanism	RMFN [47], MARN [153], MulT [126], MMUU-BA [47], RAVEN [143]	1) State-of-the-art performance on both tasks. 2) Relatively fast convergence. 3) Stabilized learning behaviour. 4) Cope with skewed and balanced datasets.
Memory Cell	MFN [152]	Capture non-aligned cross-modal interactions.
Autoencoder	MCTN [106]	1) Tackle with perturbations and missing data. 2) Fewer learning parameters.

Table 2.14: Summary of Key Findings. The first column lists the investigated key components, the second column summarizes which models are using which component, and the third column shows how different components contribute differently to solving the problem of multimodal language analysis.

These key findings were drawn from experiments over three most widely used standard benchmarking datasets in the literature, and data imbalance had been regarded as a vital issue influencing the model performance. The linguistic modality was the most informative compared to visual and acoustic modalities. We attribute that difference to the use of word embedding trained on large corpora, and not to noise issues related to the datasets. All three datasets are carefully collected, pre-processed and annotated by a world-leading group in this area, and the noise within the datasets is minimized. In practise, there is a need for investigating new approaches for training visual and acoustic embeddings. However, such an investigation is beyond the scope of this paper. Thus, we believe that our results over three high-quality and well-established large-scale benchmarking datasets can sufficiently support the conclusions.

One limitation of our study was that we used a simple approach to align modalities. Following previous work, we averaged visual and acoustic modalities throughout word intervals since advancing the SOTA was not the aim of this work. However, further investigation is needed in this direction to determine if other alignment approaches could enhance the relatively low performance of the

non-verbal modalities. In terms of the implementation, we noticed that the LSTMCell component could not speed up. That made approaches which primarily utilize recurrent cell components less efficient.

2.5 Conclusions

We have conducted a large-scale empirical comparison among SOTA approaches for multimodal human language analysis. The results show that the attention mechanism strategies, modelling pairwise modality interactions, such as MulT [126] and MMUU-BA [47], are the most effective for both sentiment analysis and emotion recognition tasks. The qualitative analysis of the study reveals that existing approaches cannot tackle ambiguous utterances, i.e., when the content of modalities is uninformative. As far as context modelling in the form of preceding utterances in conversation, current approaches either do not consider any speaker level dependency [47, 113], or support conversations with two speakers only, without can be extended to apply on multiparty conversations [59, 60], or require an artificial extension to be applied on multiparty datasets [48, 92]. The findings helped us identifying the current limitations of SOTA and provided us helpful insights to the development of more effective modality fusion models.

Chapter 3

An Overview of Applying Quantum Theory in Human Language Analysis

In Chapter 2, the empirical analysis of multimodal fusion strategies showed that attention mechanism-based models are capable of achieving unprecedented levels of performance. However, most of them cannot overcome the "black-box problem", leading to a limited utility in real-life applications. The reason is that current strategies are often designed with only performance as their design target, thus leaving aside other important aspects such as transparency, confidence, fairness or accountability [9]. To this end, researchers shall borrow concepts from cognitive science to yield a spectrum of methodological tools that can provide explainable decisions. A step forward beyond the explainable decisions and high performance, when deploying AI models, should be the consideration of how humans understand and reason about the cognitive states under which the decisions are made. Indeed, the qualitative analysis in Chapter 2 revealed that existing approaches cannot tackle ambiguous utterances, i.e., when the content of modalities is uninformative.

By contrast, quantum probability theory, as a general framework, has been extensively studied in the domain of human cognition and decision making [25]. Moreover, recent advances in quantum probabilistic neural models have achieved comparable performance for various NLP tasks [78, 139, 163], yet with better transparency due to the mapping to the well-established quantum physics meanings, and in some cases, an increased level of interpretability. In this chapter, we present the related work that exploits the mathematical formalism of quantum mechanics to model human language and establish the foundation of QT.

3.1 Related Work

3.1.1 Quantum cognition in human language

Research has found a whole range of human judgements that deviates substantially from what would be considered normatively correct according to logic or probability theory [4, 66, 128, 129]. Current generative learning algorithms are based on probability theory which cannot easily accommodate sub-optimal or irrational human decisions. When dealing with preferences under uncertainty, it seems that models based on normative theories of rational choice tend to tell how individuals must choose, instead of showing how individuals actually choose [88]. A simple solution to this problem would be to use rule-based expert systems and incorporate these cognitive biases and paradoxical decisions as rules. However, that would require an apriori understanding and identification of the cognitive biases involved, and would also lead to a set of rules growing largely with the size of the data, which is impractical. An alternative would be to rethink the fundamental core concepts that underpin decision making. If under uncertainty, humans make decisions that violate the laws of probability theory, then perhaps these laws are too limited to express human decision-making under uncertainty fully. A more general and flexible probability theory providing better insights and accommodating several paradoxical findings reported in the literature without the need to know apriori is quantum probability theory [25].

Previous studies have shown that human language understanding [19, 140] exhibits certain non-classical phenomena (i.e. quantum-like phenomena), e.g., semantic contradictions and ambiguities [18]. For example, the utterance *"two cars were reported stolen by the Groveton police yesterday"* can be interpreted in different ways, that is, the police either reported or stole the two cars [7]. That is, in some cases, human language might convey such complex characteristics, which may not fully be captured by traditional representation learning methods, which obey the classical (Kolmogorov) axioms of probability [68] and utility theory [97]. By contrast, quantum-probability based models can serve as a promising method to formulate the quantum-like phenomena and to better capture different levels and aspects of semantic units, which would be of considerable benefit for understanding human language. In particular, quantum-probability based strategies have been investigated to model combinations of words and their associations [2, 21, 22, 44, 136, 145]. The underlying idea is that the meaning of a concept, determined by the context in which it occurs, is analogous to the state of a quantum particle which is influenced by the measurement context. The mathematical formalism of quantum mechanics has also been exploited to disambiguate polysemous words [145, 146]. Finally, other work utilized QT to construct semantic spaces [3, 20, 23, 87]. Although such approaches obey a more general and flexible probability theory, they are underperformed by SOTA models. One possible reason is that the quantum probabilistic models did not encapsulate the strengths of quantum theory into data-driven approaches. Another reason might be the lack of complex numbers since non-classical phenomena can be modelled in the complex field [1, 94]. In the

next two subsections, we review initial work in NLP and IR, which integrates quantum-driven and data-driven approaches into computational models.

3.1.2 Quantum-inspired representation learning

QT is not only a physical theory but also a framework in which theories can be developed. Indeed, QT has increasingly been deployed outside physics. The application of quantum theory in representation learning began after van Rijsbergen's pioneering work [134] by integrating geometric spaces, probabilistic spaces, and logic into a unified theoretical framework. Then, the probabilistic framework of quantum theory was exploited for IR, NLP, and multimodal representation learning tasks [132]. A preliminary discussion about quantum probability formalism with application to IR can be found in [109]. Later quantum probabilities were exploited for query expansion [74, 161] and ranking [166] in IR scenarios. Crucially, quantum formalism was successfully utilised for modelling word dependencies through density matrices [122] and formulating the semantic composition of words [122] in IR tasks. Quantum-inspired models were also introduced to address NLP tasks. Early work simulated the quantum language model [122] into the neural network modelling paradigm for a question-answering task [162]. Nevertheless, it utilised real numbers only and the quantum-measurement approximated via a convolutional neural module. Recent work simulated the quantum measurement principle into an end-to-end complex-valued neural network leading to improved performance and better interpretability [78, 139].

3.1.3 Quantum-inspired multimodal representation learning

Quantum-inspired strategies were also investigated to multimodal representation learning. Wang et al. [142] proposed a tensor-based representation for an image-text IR task. They exploited statistics to capture correlations between image and text, yet without leading to performance improvement. In [164], authors derived inspiration from quantum-interference to address the decision level fusion in a heuristic manner. Even though the framework outperformed various baseline approaches, the model vaguely borrowed detached quantum concepts at different stages, implemented by real-valued components, and supported only two modalities. That is, it cannot support tri-modals (all three modalities). A recent quantum-inspired framework for conversational emotion recognition [163] derived inspiration from the concept of *weak measurement* in QT to model influences of speaker in a conversation, and the concept in [164] to fuse modalities, yet without addressing the challenge of modelling tri-modals.

3.2 Preliminaries on Quantum Theory

This section presents the fundamental concepts of quantum theory [25, 94, 99] that we exploit to construct the quantum probabilistic models. Consistent with quantum theory, we adopt the widely-used *Dirac Notations*, known as "bra-ket" notation. A complex-valued *unit* vector $\vec{u}$ and its conjugate transpose $\vec{u}^{*T}$ are denoted as a *ket* $|u\rangle$ and a *bra* $\langle u|$, respectively. The inner product of two vectors $|u\rangle$ and $|v\rangle$ is defined by $\langle u|v\rangle$, while $|u\rangle\langle u|$ and $|v\rangle\langle v|$ define operators.

3.2.1 Hilbert Space

The starting point to modelling quantum states is a set of basis states. A basis is a set of n mutually orthogonal vectors $\{|e_j\rangle\}_{j=1}^{n}$ of unit length. The vector space employed in QT is a vector space over complex numbers, called Hilbert space $\mathbb{H}$, offering the structure of an inner product to enable the measurement of angles and lengths [58, 64]. Hilbert spaces are sufficient for *completeness* because a state evolution always yield a valid state, *finiteness*, which implies completeness and is natural in Computer Science, and *complex field* because quantum operators create superposition that can only be modelled in the complex field [1, 94]

The choice of an appropriate basis depends to a large degree on how context is to be brought into the picture. For instance, if we want to estimate the conditional probability of an emotion concept among a set of basis states $\{e_j\}$ given the state s, denoted as $p(s|e_j)$, the basis shall take the form $\{e_j\}$, where each e_j represents an abstract emotion concept. Thus, in quantum theory, context is represented with the choice of basis. Note the different sets of orthonormal basis states can represent the same Hilbert space.

3.2.2 Quantum State

Any *pure state* $|s\rangle$ of a quantum system is regarded as a linear *superposition*, i.e., an appropriate weighted sum, of one set of n basis states, represented by a unit vector in $\mathbb{H}^n$. That is, the pure state $|s\rangle$ can be written as a probability distribution of complex probability amplitudes in the context of all of its associates, as follows:

$$|s\rangle = \sum_{j=1}^{n} \sqrt{r_j} e^{i\phi_j} |e_j\rangle , \tag{3.1}$$

where $\sqrt{r_j}e^{i\phi_j}$ correspond to complex probability amplitudes, r_j are non-negative scalars $\in \mathbb{R}$ satisfying $\sum_{j=1}^{n} |\sqrt{r_j}|^2 = 1$, i is imaginary number satisfying $i^2 = -1$, and ϕ_j are phases $\in [0, 2\pi]$. Before the measurement of a superposed state, there is uncertainty in that the state has no known value. That is, the state $|s\rangle$ is in a superposition of all basis states.

3.2.3 Mixed Systems

A mixed system describes the overall state of a set of pure states in probabilistic distribution. The mathematical representation of the mixed system state is a *density matrix*, which is a positive semi-definite square matrix with a unitary trace. Essentially, for a set of pure states $\{|s_i\rangle\}_{i=1}^n$ with probability weights $\{P_i\}_{i=1}^n$, the density matrix or the mixed state is computed by

$$\rho = \sum_{i=1}^{n} P_i \left|s_i\right\rangle \left\langle s_i\right| . \tag{3.2}$$

Since $\{P_i\}_{i=1}^n$ are non-negative values that sum up to 1, the complex valued density matrix ρ produced by Eq. 3.2 is always positive semi-definite with unit trace, i.e., $\rho = \rho^{*T}, tr(\rho) = 1$. The diagonal elements of ρ are always non-negative real values that sum up to 1, while the off-diagonal entries are generally complex values, i.e., *quantum interference* terms. Concretely, these off-diagonal elements are able to obtain quantum interference effects among basis states, which are non-linear functions that can be mapped to the non-linear activation functions in neural networks [55].

It is also worth noting that a pure state can be converted to a density matrix. For instance, the pure state of Eq. 3.1 can be regarded as $\rho = |s\rangle \langle s|$. Hence, the density matrix ρ can be used to represent both a pure and a mixed state in a single Hilbert Space.

3.2.4 Observable and Quantum Measurements

A measurement basis, called *observable* $\hat{O}$, establishes the environment for removing a quantum state from uncertainty. In particular, an observable $\hat{O}$ is made of a set of eigenvectors $\{|\lambda_j\rangle\}$, which forms a complete orthogonal basis of the Hilbert Space $\mathbb{H}$. Mathematically speaking, an observable $\hat{O}$ is a self-joint matrix, i.e., $\hat{O} = \hat{O}^{*T}$. *Projected-Valued Measure (PVM)* with respect to observable $\hat{O}$ yields one out of the observable eigenvalues $\{\lambda_j\} \in \mathbb{R}$ and causes the *collapse* of the state to the corresponding eigenvector. Since the quantum state collapses onto a certain eigenstate $|\lambda_j\rangle$ after measurement, applying the same observable onto the post-measurement state will always lead to the same eigenstate $|\lambda_j\rangle$. However, if the same observable $\hat{O}$ is applied to infinite copies of a quantum state, the observed values will then submit to a classical probability distribution $\{P_i\}$.

In practice, the probability of a given outcome is obtained via the projection postulate. That is, according to the Born's rule [64], the probability of the pure state $|s\rangle$ to collapse onto the eigenstate $|\lambda_j\rangle$ equals the projection of the complex-valued unit length vector $|s\rangle$ onto the basis state $|\lambda_j\rangle$. Mathematically speaking, the probability of $|s\rangle$ to collapse onto the $|\lambda_j\rangle$ is calculated by the inner product of the two vectors as follows:

$$P(s|\lambda_j) = |\langle \lambda_j | s \rangle|^2, \tag{3.3}$$

where the inner product $\langle \lambda_j | s \rangle$ is calculated as

$$\langle \lambda_j | s \rangle = |\lambda_j\rangle^{*T} |s\rangle \tag{3.4}$$

Similarly, for a mixed quantum system represented by a density matrix ρ, according to Born's rule, the probability is calculated by

$$P(\rho|\lambda_j) = tr(\rho |\lambda_j\rangle \langle\lambda_j|) = \langle \lambda_j | \rho | \lambda_j \rangle. \tag{3.5}$$

Since $|\lambda_j\rangle \langle\lambda_j|$ are a complete orthogonal basis, the resulting probabilities form a classical probability distribution summing up to 1.

However, PVMs on sub-systems of larger systems cannot be described by a PVM acting on the sub-system itself. Positive Operator-valued Measure (POVM) overcomes this constraint, by associating a positive probability for each measurement outcome, ignoring the post-measure state [100]. That is, POVM is a generalization of PVM, providing mixed information of a state for the entire ensemble of sub-systems. Mathematically speaking, a POVM M is a set of *Hermitian positive semi-definite* operators $\{E_i\}$ on a Hilbert space $\mathcal{H}$ that sum to the identity operator, i.e., $\sum_i E_i = \mathbb{1}$. For a generic pure state's density matrix ρ, where $\rho = |s\rangle \langle s|$, the probability with respect to E_i is computed as

$$P(\rho|E_i) = Tr(E_i \rho) = \langle s | E_i | s \rangle , \tag{3.6}$$

and $\sum P(\rho|E_i) = 1$. Crucially, in contrast to PVM, the elements of a POVM are not necessarily orthogonal. Practically, this means that we can skip unitary algorithms to train the proposed quantum probabilistic computational models.

3.2.5 Quantum Composite Systems

In QT, a composite system describes a compound quantum system composed of multiple individual quantum systems. That is, instead describing individual quantum systems by separate quantum states, one for each system, we are able to describe the individual states using one composite state $|s_c\rangle$. The Schmidt decomposition [99] is a particular way to express an arbitrary quantum state of a composite system. For instance, suppose A and B are two n_A-dimensional and n_B-dimensional spaces, respectively. Then, any state vector $s_c \in A \otimes B$ can be expressed as a linear combination of an arbitrary basis of the product space $|i\rangle \otimes |j\rangle$, where $|i\rangle$ and $|j\rangle$ are two arbitrary bases of A and B, respectively. That is,

$$|s_c\rangle = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \sqrt{c_{i,j}} |i\rangle \otimes |j\rangle , \tag{3.7}$$

where $\sqrt{c_{i,j}}$ are probability amplitudes satisfying $\sum_{i,j} c_{i,j} = 1$. Then, according to Schmidt decomposition theorem, the state s_c can be expressed as a superposition of the product basis states, that is,

$$|s_c\rangle = \sum_{i=1}^{min(n_A,n_B)} \sqrt{\gamma_{i,i}}\,|i_A\rangle \otimes |i_B\rangle\,, \tag{3.8}$$

where $|s_c\rangle \in A \otimes B$, $|i_A\rangle \in A$ and $|i_B\rangle \in B$ are orthonormal bases for $\mathbb{H}^{n_A}$ and $\mathbb{H}^{n_B}$, respectively, and $\gamma_{i,i}$ are real and non-negative coefficients satisfying $\sum_{i=1}^{min(n_A,n_B)} \gamma_{i,i} = 1$. Note that this theorem does not generalize to three or more spaces, therefore, can be used for bipartite quantum states only.

3.2.6 Reduced Density Matrix

Reduced density matrix is used to construct representations of sub-systems from a composite quantum system [100]. Suppose a state $\rho \in \mathcal{H}_{AB}$ for a bi-particle system composed of sub-systems A and B. The density matrix ρ_A of the sub-system A so that applying any measurement $M \in \mathcal{H}_A$ onto it to yield the same result as applying the measurement $M \otimes I$ on the composite system state ρ with the sub-system B unchanged, according to Eq. 3.5, is given by

$$tr(M\rho_A) = tr((M \otimes I_B)\rho), \forall M \in \mathcal{H}_A \tag{3.9}$$

The solution ρ_A of the above equation is obtained by taking the *partial trace* of ρ over the sub-system B:

$$\rho^A = tr_B(\rho), \tag{3.10}$$

where the partial trace is defined as follows: suppose ρ can be expressed as $\rho = \sum_{ijkl} c_{ijkl}\,|e_i\rangle_A \langle e_j| \otimes |f_k\rangle_B \langle f_l|$, then $\rho_A = tr_B(\rho) = \sum_{ijkl} c_{ijkl} \langle f_l|f_k\rangle |e_i\rangle \langle e_j|$. In the matrix form, this stands for computing traces for all blocks corresponding to each sub-system division. Fig. 3.1 shows the case of taking the partial trace over a two-qubit system state. Since the partial trace operation obeys commutative law, the reduced density matrix can be properly defined over any subset of a composite system of an arbitrary scale (i.e., number of systems).

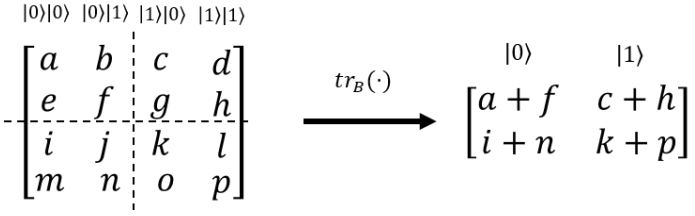


Figure 3.1: Illustration of partial trace. The left-hand side is a density matrix of a two-qubit system. The partial trace is performed over system B. The right hand side shows the resulting 2 by 2 reduced density matrix of system A.

3.2.7 Quantum Entanglement

In QT, the way that quantum systems can be correlated is fundamentally different from the classical theory. When a composite quantum system evolves under a Hamiltonian [1] that includes interactions between individual sub-systems, the resulting state of the composite system is, in general, no longer *separable*. The fact that individual quantum systems are *non-separable*, i.e., cannot be expressed as a tensor product of individual systems (see Eq. 3.12), make them *entangled* [64, 93]. This implies that the joint probability distribution of the composite system would no longer factor in the respective probabilities of individual sub-systems. Entanglement is beyond the classical correlations, and violates the classical theories, such as Bell inequalities [71].

The bipartite Von Neumann *entanglement entropy* S [138] is a measurement of the degree of quantum entanglement for a composite pure state. For instance, for an arbitrary bipartite quantum state consisting of two substates, i.e., $A \in \mathbb{H}^{n_A}$ and $B \in \mathbb{H}^{n_B}$, the Von Neumann entropy is calculated as follows,

$$S = S_A = S_B = -\sum_{i}^{min(\mathbb{H}^{n_A}, \mathbb{H}^{n_B})} |a_i| log(|a_i|), \tag{3.11}$$

where a_i are singular values of the Schmidt decomposition of the bipartite quantum state over either the A or B individual system. Eq. 3.11 makes clear that the entanglement entropy is the same regardless of whether one decomposes over the A or B sub-system. Crucially, if the entropy S is zero, there is no entanglement and $|s_c\rangle$ is separable into $|s_A\rangle$ and $|s_B\rangle$, that is

$$|s_c\rangle = |s_A\rangle \otimes |s_B\rangle \,. \tag{3.12}$$

Note that Von Neumann entropy's reciprocal value is give by the so-called Schmidt number

$$K = \frac{1}{\sum_{i=1}^{min(\mathbb{H}^{n_A}, \mathbb{H}^{n_B})} a_i^2}, \tag{3.13}$$

defined as the effective number of non-zero coefficients in the Schmidt decomposition [103], i.e., the number of effective degrees of freedom contributing to the entanglement. The Schmidt numbern K is defined on the interval $[1, min(\mathbb{H}^{n_A}, \mathbb{H}^{n_B})]$ and $K \sim min(\mathbb{H}^{n_A}, \mathbb{H}^{n_B})$ corresponds to maximal entanglement. If K equals 1, then the composite state is a separable state.

3.2.8 Quantum Dynamics

In quantum mechanics, there is a law under which quantum systems change over time. Such a law should be consistent with the superposition principle and the probabilistic interpretation. In particular, the state at the current time t should be a linear operator acting on the state at the

[1] A Hermitian linear operator that can be derived from the Euler–Lagrange equations of motion of the associated "classical" system.

previous time $t-1$. Mathematically speaking, the motion of the state $|s\rangle$ is given by the Schrödinger equation [94] as follows:

$$i\hbar\frac{\partial}{\partial t}|s(t)\rangle = \hat{H}|s(t)\rangle\,, \tag{3.14}$$

where $\hat{H}$ is Hamiltonian. If $\hat{H}$ does not change in time, then the time evolution operator for the time t is the unitary U operator. Hence, the state evolves according to

$$|s(t)\rangle = U|s(t-1)\rangle\,, \tag{3.15}$$

3.2.9 Summary

We now have a firm notion of how to describe the states of quantum systems, what quantum measurement is about, and how quantum systems change with time. Additionally, complex-valued representations can model non-linear interactions among basis states, corresponding to context. On the other hand, entanglement, being beyond classical correlations, can be exploited to model correlations across distinct modalities. Finally, the Hamiltonian evolution is another way to model time-series information. It remains an open problem to elaborate on how the axioms of quantum theory can be injected into various components, and then, how we can join them into quantum probabilistic computational models, encapsulating the strengths of quantum theory-driven approaches into a data-driven approach. This dissertation aims to make a step forward to address the problem.

3.3 Conclusions

Quantum probability theory is a more general and flexible probability theory, providing better insights and accommodating several paradoxical findings reported in the literature without the need to know apriori. On the other hand, recent advances in quantum probabilistic neural models achieved comparable performance to the SOTA [78, 139, 162, 163, 164], yet with better transparency due to the mapping to quantum physics meanings, and in some cases, an increased level of interpretability [78]. However, existing work utilized either real numbers only [162, 163, 164] or modelled quantum states as a classical mixture [78, 139], which cannot fully exploit the potential of quantum probabilistic description. Moreover, quantum probability formalism has been exploited to model bi-modals only (i.e., any of two modalities), borrowing vaguely detached quantum concepts at different stages,[163, 164]. It is an open research question of how the axioms of QT shall be injected into unified computational models to address fusion of multiple modalities (e.g., more than any two modalities), encapsulating the strengths of quantum theory-driven approaches into data-driven approaches. This dissertation aims to make a step forward to address the problem.

Chapter 4

A Quantum-inspired Multimodal Fusion Framework

In this chapter, we tackle the crucial challenge of fusing different modalities of features for utterance-level multimodal sentiment analysis. Existing neural network approaches largely model multimodal interactions in an implicit and hard-to-understand manner. We address this limitation with inspirations from quantum theory, which contains principled methods for modelling complicated interactions across distinct modalities. In particular, we draw the analogy of the electron property states in terms of the multimodal fusion of input features, by simulating the quantum measurement postulate and its procedural steps for video sentiment analysis.

In our quantum-inspired multimodal fusion (QMF) framework, the word interaction within a single modality and the interaction across modalities are formulated as superposed and decomposable composite states, respectively, at different stages. Sentiment decisions are made via the concept of *quantum measurement*, which is a natural choice given the quantum state representation of multimodal sentences. Concretely, an *observable* is introduced to measure the probabilities of the multimodal sentences in the states of main sentiment-related aspects. The probability values are then passed to a fully connected layer to predict the final sentiment. The complex-valued neural network implementation of the framework achieves comparable results to SOTA neural models on two benchmarking video sentiment analysis datasets and high transparency due to the mapping to quantum physics meanings. In the meantime, we produce the uni-modal and bi-modal sentiment directly from the model to interpret the final multimodal decision.

4.1 Introduction

As discussed in Chapter 2, in most of three directions of performing multimodal fusion, the way the modalities interacted was often vague and implicit for intra-modal, and inter-modal interactions [13]. This phenomenon is closely related to the interpretability issue and the broader concept of XAI. Interpretability has become a significant concern for machine learning models. As those models have brought about remarkable performance boosts, researchers are looking for ways to understand the model, in order to know whether we can trust it and deploy it in real work [82] or whether it contains privacy or security issues [63]. Existing models in multimodal sentiment analysis heavily rely on neural structures to fuse multimodal data, which often behave like black-boxes with few numerical constraints and purely data-driven assignment. As a result, these models invariably suffer from low interpretability.

In this study, we investigate a quantum-inspired approach for fusing multimodal data in an attempt to provide a principled view of multimodal fusion from a quantum perspective. The inspiration stems from the manifestation of non-classical phenomena in human cognition and decision, which violates classical probability theory but adopts a compact explanation via QT [25]. QT has stimulated the successful construction of quantum-inspired models for human cognition-related tasks, such as information retrieval (IR) [77, 122] and language understanding [78, 139]. As a typical human cognitive task, however, multimodal sentiment analysis has received little attention from a quantum-inspired viewpoint [164], due to the challenge in modelling complicated interactions across different modalities in a quantum manner. In this work, we take a step forward and present a novel quantum-theoretic multimodal fusion framework.

4.2 Model

We now present the quantum-inspired framework for multimodal sentiment analysis. Since Hilbert Space is the mathematical foundation of any quantum-theoretical framework, it is necessary to define the Hilbert Space. In the remaining part of the section, we define the Hilbert Space grounding the proposed framework and introduce the formulation of words, sentences, and sentiment decisions. The task is an utterance-level sentiment analysis as discussed in Chapter 2, Section 2.2.1.

4.2.1 Multimodal Hilbert Space

We generally view a multimodal sentence as a composite quantum system of individual modalities. Hence, in our framework, the Hilbert Space is a composition of uni-modal Hilbert Spaces for single modalities, referred to as *Multimodal Hilbert Space* $\mathbb{H}_{mm}$. In multimodal sentiment analysis, we focus exclusively on the textual, visual, and acoustic modalities. However, it is worth noting that our framework is general and could be adapted to any number of modalities.

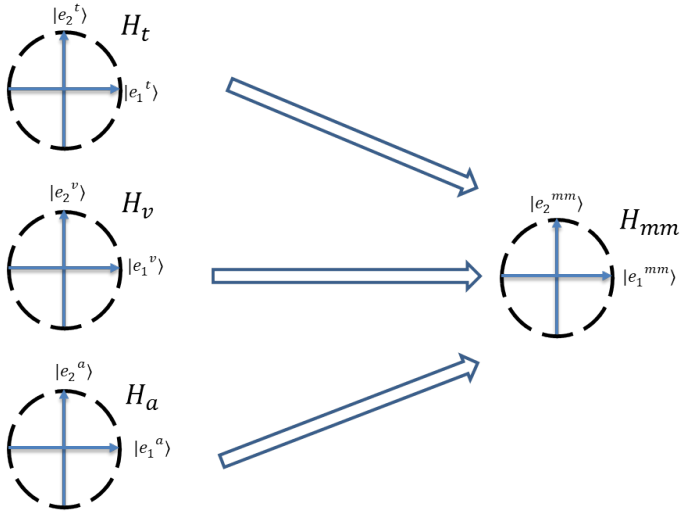


Figure 4.1: The Multimodal Hilbert Space $\mathbb{H}_{mm}$ composed of textual, visual and acoustic Hilbert Space $\mathbb{H}_t, \mathbb{H}_v, \mathbb{H}_a$. $|e_j^t\rangle, |e_j^v\rangle, |e_j^a\rangle, |e_j^{mm}\rangle$ denotes a basis state of $\mathbb{H}_t, \mathbb{H}_v, \mathbb{H}_a, \mathbb{H}_{mm}$ respectively.

Suppose $\mathbb{H}_t, \mathbb{H}_v, \mathbb{H}_a$ denote the Hilbert Space for textual, visual, and acoustic modalities spanned by the basis states $\{|e_i^t\rangle\}_{i=1}^{t_{dim}}$, $\{|e_j^v\rangle\}_{j=1}^{v_{dim}}$ and $\{|e_k^a\rangle\}_{k=1}^{a_{dim}}$, respectively. $\mathbb{H}_{mm}$ is then expressed as $\mathbb{H}_{mm} = \mathbb{H}_t \otimes \mathbb{H}_v \otimes \mathbb{H}_t$ with a set of basis states $\{|e_i^t\rangle \otimes |e_j^v\rangle \otimes |e_k^a\rangle\}_{i=1j=1k=1}^{t_{dim}v_{dim}a_{dim}}$. The basis can be re-written as $\{|e_l^{mm}\rangle\}_{l=1}^{t_{dim}\times v_{dim}\times a_{dim}}$ for simplification purposes, where each $|e_l^{mm}\rangle$ is a tensor product of $|e_i^t\rangle$, $|e_j^v\rangle$, $|e_k^a\rangle$ for some i, j, k. Fig. 4.1 shows the multimodal Hilbert Space in the composition of three individual Hilbert Spaces.

4.2.2 Word State

A word w is formulated as a pure state $|w\rangle$ on $\mathbb{H}_{mm}$. Since a word is associated with a textual, visual and acoustic feature vector, we are able to construct its uni-modal state representation $|w^t\rangle$, $|w^v\rangle$ and $|w^a\rangle$ in $\mathbb{H}_t$, $\mathbb{H}_v$, $\mathbb{H}_a$, respectively. It is then an open issue to construct $|w\rangle$ based on the respective uni-modal states of w. In this work, we assume $|w\rangle$ to be a product state of uni-modal states, i.e., $|w\rangle = |w^t\rangle \otimes |w^v\rangle \otimes |w^a\rangle$, as shown in Fig. 4.2. This simple strategy is employed for the reasons below:

- By implementation, it gives rise to a tensor-based fusion of multimodal signals, which is believed to be a meaningful and useful approach to capture inter-modal interactions [15, 85, 151]. In particular, it explicitly aggregates features of three modalities by means of multiplication, while other models [47, 126, 143, 152, 153] instead rely on additional structures to fuse uni-modal features in a more implicit manner.

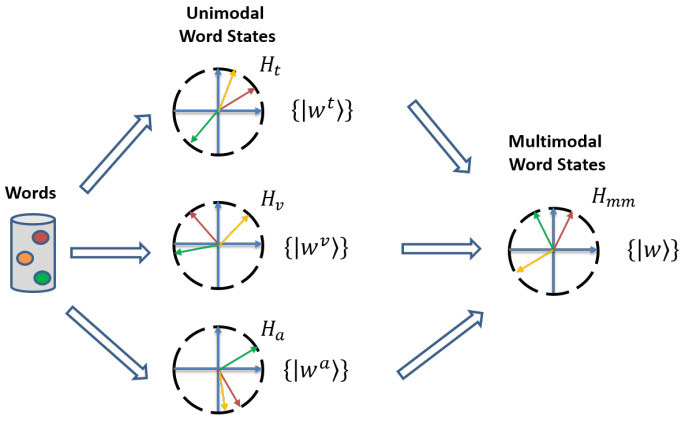


Figure 4.2: Multimodal Word Representation. Each color indicates one word. The multimodal word state $|w\rangle$ for word w is a tensor product of its uni-modal states $|w^t\rangle$, $|w^v\rangle$ and $|w^a\rangle$.

- When uttering one word or one sentence, a person may aim at expressing different sentiments under different situations. A single word has different multimodal representations under different visual-acoustic contexts based on word-dependent textual representation and word-independent visual and acoustic representations. As a result, different sentiments of a specific word or sentence can be accounted for by this multimodal word representation.

4.2.3 Sentence State

We formulate a sentence as a mixture of individual word states $\{|w\rangle\}$ in the sentence. The mixed state $\rho \in \mathbb{H}_{mm}$ of a sentence is produced by the individual word states in the sentence in the form of a weighted quadratic summation:

$$\rho = \sum_i \lambda_i \left|w_i\right\rangle \left\langle w_i\right| \tag{4.1}$$

where $\{\lambda_i\}$ are convex coefficients, i.e., $\sum_i \lambda_i = 1$ in order to guarantee $Tr(\rho) = 1$. λ_i is a word-dependent weight that reflects the importance of the word w_i in the sentence. The sentence mixed state ρ is visualized as an ellipse constructed by unit vectors of words in the sentence in Fig. 4.3. The ellipse representation is due to the fact that a density matrix assigns a probability measure on the Hilbert Space from the quantum probability point of view. Please refer to [122] for a detailed explanation.

Even though a density matrix is constructed from a particular set of word weights, it corresponds to many possible mixture weights of the same set of words (Section 3.2.3). As a result, it is capable of formulating different word combinations under different contexts. As a probability measure on the Multimodal Hilbert Space, the density matrix is a sentence representation in terms of uni-modal

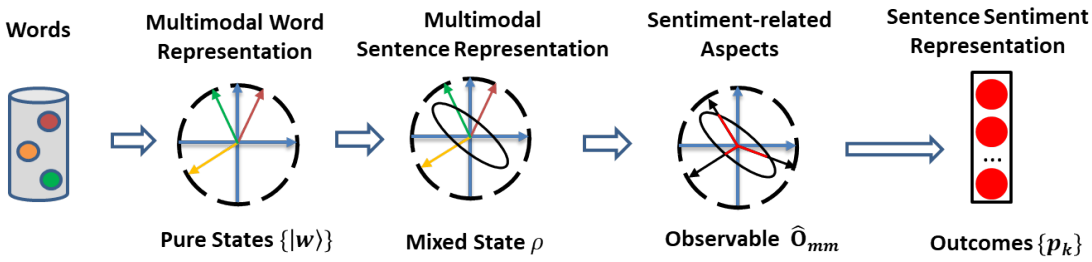


Figure 4.3: Our framework. Each colored ball indicates a word in the multimodal sentence, represented as a unit vector of the same color in the Multimodal Hilbert Space. The sentence is represented by a mixed state visualized as a black ellipse. The eigenstates of the observable are unit vectors in black color. The squared length of the intersection between each unit vector and the ellipse (in red) is the measurement probability for the respective eigenstate. The sentence sentiment representation is composed of all probability values represented by red balls.

features in combination from a classical perspective. The representation is a separable state rather than an entangled state because ρ can be re-written as

$$\begin{aligned} \rho &= \sum_i \lambda_i (|w_i^t\rangle \otimes |w_i^v\rangle \otimes |w_i^a\rangle)(\langle w_i^t| \otimes \langle w_i^v| \otimes \langle w_i^a|) \\ &= \sum_i \lambda_i (|w_i^t\rangle \langle w_i^t|) \otimes (|w_i^v\rangle \langle w_i^v|) \otimes (|w_i^a\rangle \langle w_i^a|) \\ &= \sum_i \lambda_i \rho_i^t \otimes \rho_i^v \otimes \rho_i^a, \end{aligned} \tag{4.2}$$

where $\rho_i^m = |w_i^m\rangle \langle w_i^m| \in \mathbb{H}_m$ for $m \in \{t, v, a\}$. From Section 3.2.5 and 3.2.7, ρ is separable with respect to the three uni-modal Hilbert Spaces by definition. Consequently, the framework considers word-level interactions via the concepts of mixture and superposition on the feature level, while the interactions across different modalities are largely absent from the feature level. Instead, the inter-modal interactions are implemented in the sentiment decision process, as outlined in the next paragraphs.

4.2.4 Sentiment Measurement

Based on the multimodal sentence representation, a component is needed to operationalize the sentiment judgment process. To this aim, we link sentiment judgment to quantum measurement and "measure" the "sentiment state" of a multimodal sentence.

We hypothesize that there are K sentiment-related aspects or topics, such as the aspects of aspect-based sentiment analysis. The sentence will collapse onto one of them after the measurement. The probabilities over the aspects after repeating the measurements can be seen as a sentiment

characterization of a multimodal sentence, which could be used to determine the sentence sentiment.

Mathematically, the multimodal observable $\hat{O}_{mm}$ is associated with a set of aspect ids $\{k\}_{k=1}^{K}$ as eigenvalues and a set of aspect representations $\{|v_k\rangle\}_{k=1}^{K}$ as eigenstates. Hence $\hat{O}_{mm}$ can be expressed as

$$\hat{O}_{mm} = \sum_{k} k \, |v_k\rangle \, \langle v_k| \, . \tag{4.3}$$

After the experiment, the multimodal sentence ρ will collapse onto the k-th aspect at a likelihood of P_k:

$$P_k = \langle v_k| \, \rho \, |v_k\rangle \tag{4.4}$$

The final sentiment judgment is given based on the clues from each sentiment-related aspect, and the probability values $\{P_k\}_{k=1}^{K}$ are taken to generate the sentence sentiment. In Fig. 4.3, a set of unit-norm vectors is associated with the observable $\hat{O}_{mm}$. The squared lengths of their intersections with the density matrix ellipse ρ are the measurement probabilities $\{P_k\}_{k=1}^{K}$ that reflect sentence sentiment.

It is worth noting that each sentiment-related aspect is a pure state $|v_k\rangle$, which is always a composite state of the three uni-modal systems. Hence, the aspects are abstract concepts over the whole multimodal space that can hardly be mapped to human-understandable notions. Instead, each aspect can be seen as a composite multimodal sentiment decision of uni-modal sentiment decisions. The observable $\hat{O}_{mm}$ is uniquely represented by the eigenstates $\{|v_k\rangle\}_{k=1}^{K}$. In the rest of the chapter, we use $\{|v_k\rangle\}_{k=1}^{K}$ and $\hat{O}_{mm}$ interchangeably to represent an observable.

The way uni-modal sentiment decisions are aggregated can be displayed with the help of a reduced density matrix. The reduced density matrix allows us to obtain the statistically equivalent observable for bi-modal and uni-modal systems so that the decisions entailed in the tri-modal system can be inferred by applying the observable onto the respective sentence representation. The details of this process are introduced in Section 4.4.4.

4.3 Methodology

This section outlines the neural network implementation of our quantum-inspired multimodal fusion framework for multimodal sentiment analysis. Complex values are pivotal to the formulation of quantum concepts, so our network is composed of complex-valued units as an authentic formulation of the quantum-inspired multimodal fusion process. Fig. 4.4 shows the architecture of the network. Next, we introduce the way to handle complex values for each network component, so that the network weights could be learned in the same way as any classical neural network.

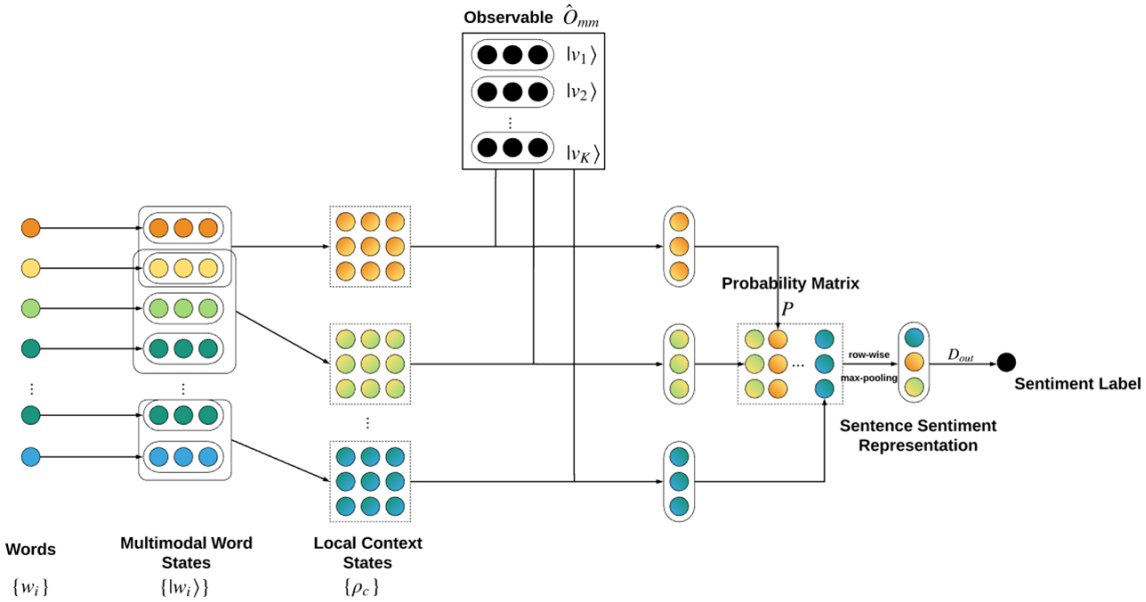


Figure 4.4: The quantum-inspired multimodal fusion network. The multimodal word states are obtained via complex-valued multimodal word embedding. The local context states are constructed from individual word states under the global weighting and local mixture strategy. The multimodal observable is applied to each context state in the measurement step, and the obtained probability matrix is row-wise max-pooled and passed to a neural network to produce the final sentiment.

4.3.1 Complex-valued Multimodal Word Embedding

As previously introduced, the multimodal word state is $|w_i\rangle = |w_i^t\rangle \otimes |w_i^v\rangle \otimes |w_i^a\rangle$ in the Multimodal Hilbert Space. The task is to map real-valued input features to complex-valued unit vectors for each word under each modality. To this aim, we adopt the modulus-argument form for a complex number. Each uni-modal state for a word w is represented as

$$\begin{aligned} |w\rangle &= r_1 e^{i\theta_1} |e_1\rangle + ... + r_n e^{i\theta_n} |e_n\rangle \\ &= [r_1 e^{i\theta_1}, ..., r_n e^{i\theta_n}], \end{aligned} \tag{4.5}$$

where i is the imaginary number satisfying $i^2 = -1$, the moduli $R = [r_1, ..., r_n]$ form a real unit vector, and the arguments $\Theta = \{\theta_1, ..., \theta_n\}$ are in $[-\pi, \pi]$ each. In the modulus-argument form, any operation on the complex numbers will lead to a non-linear combination of the constituent moduli and arguments. If the moduli and arguments could be appropriately assigned with different features, a non-linear feature combination is naturally produced.

Different policies are employed to assign the moduli and arguments from the input features for different modalities. The textual modality possesses a word-dependent distributed representation,

while the features for non-textual modalities are word-independent and non-trainable. In contrast to the textual modality that each word is represented as word embedding, visual and acoustic embeddings are extracted in the pre-processing step. This implies that visual and acoustic embeddings are different in different video clips for the same word whilst words in the textual representation are the same across video clips. Hence, visual and acoustic embeddings should be fixed during training, i.e., they should be not trainable. Therefore, the moduli $R^t = [r_1^t, ..., r_n^t]$ for w are constructed from the pre-trained word embedding $E(w)$ via a deep neural network D_t, while the moduli of visual and acoustic modalities R^v, R^a are obtained via deep neural networks D_v and D_a from the respective input feature vectors V_w and A_w (note that they do not depend on word w). Precisely, $R^t = N(D_t(E(w)))$, $R^v = N(D_v(V_w))$, $R^a = N(D_a(A_w))$ with $N(\cdot)$ as the vector $L2$-normalization function. In line with the Tensor Fusion Network [151], D_t is composed of an LSTM layer followed by two fully connected layers, while D_v and D_a are three stacked fully connected layers. They mainly serve as dimension reduction models, ensuring that the dimensionality of the multimodal Hilbert Space is computationally affordable. Moreover, the textual LSTM structure memorizes the sequence information, complementing the quantum-inspired framework that ignores word order. This step produces a low-dimensional representation R^t, R^v and R^a for the moduli. They are then unit-normalized to meet the unit-norm constraints.

The way to initialize arguments for each modality is as follows: for textual modality, we initialize the arguments of sentiment words regarding their respective sentiment polarity. To this goal, we use a prior polarity lexicon of approximately 155,000 words, named SentiWords [45], to derive prior polarity lexica for sentiment analysis. In particular, a positive word is initialized with a zero vector, and a negative word is initialized with a vector of π, while non-sentiment words are assigned with a vector of $\pi/2$ for their respective textual arguments. The assumption behind this is that the individual word sentiment influences the sentence sentiment, and we aim at leveraging the word sentiment by linking it with the textual arguments. Here we map an argument to the sentiment with the cosine function so that the arguments of π, $\pi/2$, 0 are mapped to -1,0,1, indicating a negative, neutral and positive sentiment respectively. Since only a rough estimation of word sentiment is present in a sentiment dictionary, it is used as initial values of the arguments subject to fine-tuning together with the other network components. For the non-textual modalities, the arguments Θ^v and Θ^a are set to be word-dependent. Even though the non-textual representation is word-independent, different representations of the same word may still share some information that possibly helps to make the sentiment judgment. Hence we build the quantum-inspired framework to learn the arguments Θ^v and Θ^a respectively based on uni-modal features. The learned arguments are used as initial values of arguments that are fine-tuned on the tri-modal data.

4.3.2 Mixture

In the previous section, we have outlined the formulation of a sentence as a mixture of individual words following Eq. 3.2. To adapt this step to the multimodal sentiment analysis scenario, one needs to answer the crucial question of determining the word-dependent probability weights $\{P_i\}$ in Eq. 3.2. Furthermore, the sentiment of the sentence is often determined by local contexts, i.e., consecutive words within local windows, rather than the whole sentence. Therefore, another issue falls on the identification of contexts that provide crucial clues to judge the sentiment.

In this work, we addressed both issues through *global weighting and local mixture*, which has been taken for constructing text-based language representation [78]. Essentially, we assigned a global weight to each word and use the global weights to determine the density matrices for local word contexts. As for the weighting scheme, the weight Λ_i of a word w_i should be composed of its weights under all three modalities. Hence, we applied a weighted sum of uni-modal weights to compute Λ_i:

$$\Lambda_i = \beta_t \Lambda_i^t + \beta_v \Lambda_i^v + \beta_a \Lambda_i^a, \tag{4.6}$$

where $\Lambda_i^t = ||D_t(E(w_i))||_2$, $\Lambda_i^v = ||D_v(V_{w_i})||_2$, $\Lambda_i^a = ||D_a(A_{w_i})||_2$ are the $L2$-norms of the contracted textual, visual and acoustic feature vectors of w_i respectively. $\{\beta_m \in [0,1], m \in \{t,v,a\}\}$ are modality-specific weights that sum up to 1.

The weighting scheme were followed by the local mixture of words in the multimodal sentence. Specifically, a set of local contexts were identified, and the words in each context c were mixed in a quantum manner (i.e., Eq. 3.2) to produce a density matrix ρ_c. The mixture weights were produced by softmax-normalizing the word weights within each context so that the outcome of the mixture is always a legal density matrix. The approach to extract local contexts from the sentence is an open issue. In this work, we applied sliding windows of varying lengths through the whole sentence, each producing a density matrix representing a local n-gram. Hence, rather than a single density matrix, a set of matrices were produced by the local mixture component. In the measurement step, the most representative contexts were identified in a data-driven fashion, as outlined in Section 4.3.3.

It is worth noting that the L2-norms of feature vectors were used to fit the construction of complex-valued word embedding. In order to ensure that each uni-modal representation can be interpreted as a pure state, vector L2-normalization was applied, and vector norms were hence discarded. The vector norm somehow reflects the semantic intensity, which may be indicative of the combination of words in a local context. From a quantum perspective, the absolute number of each pure state should be considered when mixed together.

4.3.3 Measurement

The measurement component needs to handle a set of C density matrices $\{\rho_c\}$ for local contexts, and identify the discriminating contexts to sentiment classification. To achieve this purpose, a

single observable $\{|v_k\rangle\}_{k=1}^{K}$ was performed to the set of density matrices, each generating a set of probability values via Eq. 3.5. As a result, a K-by-C matrix of probability values was produced by the measurement, each entry corresponding to the likelihood of a local context collapsing to an eigenstate. Then a row-wise maximum pooling was conducted to get the most similar local context for each of the K sentiment-related aspects. The respective probabilities, i.e., the K maximum probability values, were treated as the sentence sentiment representation. A neural network D_{out} was built on its basis to produce the final sentiment prediction of the multimodal sentence.

We aimed to learn the eigenstates or sentiment-related aspects $\{|v_k\rangle\}_{k=1}^{K}$ from the data, as it was difficult to map them to concrete notions beforehand. A deviation from the standard definition of observable was then employed: the set of eigenstates did not necessarily form an orthonormal basis of the Multimodal Hilbert Space, but were instead of a predefined number of K and not hard-coded as orthogonal to each other. The reasons were two-fold. On the one hand, different abstract sentiment-related aspects are not necessarily independent of each other in practice. On the other hand, it is not computationally affordable to ensure mutual orthogonality of measurement states during training, even though there are already algorithms for training mutually orthogonal vectors [8, 147].

4.3.4 Network Learning

The network weights include word embeddings E, arguments Θ_t, Θ_a, Θ_v, modality-specific weights $\{\beta_m \in [0,1], m \in \{t,v,a\}\}$, observable $\{|v_k\rangle\}_{k=1}^{K}$, and neural network weights D_t, D_a, D_v and D_{out}. Note that the mixture step does not contain any trainable weights.

As for the initialization, E was initialized with existing word embeddings. The textual arguments Θ_t was initialized in a word sentiment related manner as introduced in Section 4.3.1. In order to initialize the visual and acoustic arguments, we pre-trained the framework on respective uni-modal data consisting of a dimension reduction network, a global mixture of all words in the sentence, a measurement component and the output network D_{out} with random initialized argument. The eigenstates in the observable had random-initialized arguments and random-initialized unit-norm moduli.

During training, the moduli and arguments of complex-valued inputs were trained separately with unit-norm constraints imposed on the moduli part. The intermediate complex-valued layers were implemented with real and imaginary parts for inputs and outputs, in order to back-propagate the loss function to real and imaginary parts separately.

4.3.5 Network Interpretation

Our network captures multimodal interactions by borrowing concepts from quantum theory. For the quantum-like process to be understandable for human beings, we propose an approach to interpret the network. Essentially, the model captures word interactions via superposition and inter-modal

interactions through the composition of the superposed states. Both levels of interactions could be explicitly understood from the learned tri-modal model as follows:

I) The uni-modal and bi-modal decisions entailed in the learned model can be computed for a target sample. The uni-modal word states and weights can be computed from the learned model, allowing us to compute global word weights via Eq. 4.6 for any subset of the three modalities, and then the mixed state of any local context on its basis. The corresponding observable for the respective modalities is computed by taking the reduced density matrix (Section 3.2.6) of the learned eigenstates so that the measurement can be applied to the set of obtained density matrices. The probabilities are row-wise max-pooled and passed to the learned D_{out} to generate the sentiment label for the target subset of modalities.

II) The multimodal sentiment judgment for any word or word combination can be inferred from the learned model. With the learned observable and output network D_{out}, the sentiment label for any density matrix $\rho \in \mathbb{H}_{mm}$ can be produced. A word adopts a density matrix representation (Section 3.2.3). The density matrix of any combination of words, such as local contexts, can also be computed as a mixture of word states, with mixture weights being softmax-normalized global word weights. Hence we can check the sentiment for each word or word combination determined by the learned model.

The point that the learned model could be directly leveraged to generate results for part of the data is crucial to address the interpretability issue because that refers to the model's authentic behavior. When the models require re-training on the subset of data, on the other hand, the result cannot be safely interpreted as the performance of the original model anymore. In the multimodal sentiment analysis context, if a tri-modal network needs to be re-trained to predict sentiments for uni-modal or bi-modal data, it will remain doubtful whether the results could be used to "interpret" the behavior of the original tri-modal model on the uni-modal or bi-modal systems.

However, instead of directly taking the learned network to give predictions, most prior work in this field requires re-training the model based on uni-modal and bi-modal data. In particular, LSTM-based approaches involve concatenations of uni-modal hidden units to produce the inter-modal dynamics, so one cannot directly apply the component to a bi-modal and uni-modal case due to the dimension inconsistency. In MulT [126] and tensor-based approaches [15, 85, 151], the shape of multimodal sentence representation is relevant to the number of modalities, so the neural structures should be re-trained to predict sentiments based on bi-modal and uni-modal representations. MCTN [106] has a single-directional structure for the second seq-to-seq component, so it could only be used to predict part of the bi-modal and tri-modal sentiment, depending on the order of the modalities put into modelling.

To the best of our knowledge, LMF [85] is the only prior model that facilitates direct computation of uni-modal and bi-modal sentiments from the learned tri-modal network. However, an analysis of such a property is missing from the original LMF paper. This chapter identifies this property

of our model, and presents the exact prediction results of the tri-modal network on uni-modal and bi-modal data in Section 4.4.4.

4.4 Experiments

4.4.1 Experimental Setup

The experiments were conducted on CMU-MOSI [155] and CMU-MOSEI [157] SOTA benchmakring datasets for video sentiment analysis. The details about the datasets and feature extraction have been described in Section 2.2.2. However, in this work there is a crucial additional pre-processing step. In particular, we merged two version of datasets to obtain import words that there are missing in the version of CMM-Multimodal SDK. Such pre-processing step could substantially affect the overall performance, since linguistic modality is the most informative compared to visual and acoustic modalities.

To evaluate the proposed model, we conducted a comprehensive comparison with the following SOTA models introduced in Chapter 2: 1) Early-Fusion LSTM (EF-LSTM), 2) Late-Fusion LSTM (LF-LSTM), 3) Multi-Attention Recurrent Network (MARN) [153], 4) Memory Fusion Network (MFN) [152], 5) Tensor Fusion Network (TFN) [151], 6) Low-rank Multimodal Fusion (LMF) [85] and 7) Multimodal Transformer (MulT) [126], the details of which were given in Section 2.2.5. Finally, we used the same evaluation metrics as were described in Section 2.2.3, in agreement with [157].

4.4.2 Performance Analysis

The performance on CMU-MOSEI and CMU-MOSI is shown in Tables 4.1 and 4.2 respectively. The bold values refer to the highest performance out of all the models for a specific metric. For each model, the percentage difference from the best score (%Δ) is shown in parentheses next to its absolute performance. The best hyperparameters for CMU-MOSEI were $t_{dim} = v_{dim} = a_{dim} = 10$, local context length $l = \{1, 3\}$, the number of eigenstates $K = 20$, last hidden layer size $h = 48$, batch size $bs = 32$, and learning rate $lr = 0.002$. The best settings for CMU-MOSI were $t_{dim} = v_{dim} = a_{dim} = 10$, $l = \{1, 2\}$, $K = 30$, $h = 24$, $bs = 32$, and $lr = 0.001$ respectively.

Both tables indicate close results between our QMF and the best-performed models in the experiment. In particular, QMF obtained the best performance in MAE and Correlation and ranked second in binary accuracy and F1 value on CMU-MOSI. QMF was less competitive on CMU-MOSEI compared to other models, but it marginally underperformed the best model at a relative difference of less than 2.5% in all metrics. A significant performance discrepancy of over 2.5% between QMF and the best model was observed solely in 7-level accuracy on CMU-MOSI. We posit that it is because CMU-MOSI is a smaller dataset, and a minor increase in the number of wrong samples may

Model	Acc-7	Acc-2	F1	MAE	Corr
Vanilla LSTM					
EF-LSTM	0.4753 (2.86%)	0.7921 (2.43%)	0.7895 (2.01%)	0.6560 (3.88%)	0.6268 (5.22%)
LF-LSTM	0.4719 (3.56%)	0.7911 (2.55%)	0.7855 (2.50%)	0.6669 (5.61%)	0.6102 (7.72%)
LSTM+					
MARN [153]	0.4837 (1.14%)	0.8090 (0.34%)	0.8014 (0.53%)	**0.6310**	0.6515 (1.48%)
MFN [152]	0.4448 (9.09%)	0.8031 (1.07%)	0.7925 (1.64%)	0.7044 (11.54%)	0.6562 (0.77%)
Tensor					
TFN [151]	**0.4893**	**0.8118**	**0.8079**	0.6465 (2.38%)	0.6515 (1.48%)
LMF [85]	0.4824 (1.41%)	0.8064 (0.67%)	0.8057 (0.27%)	0.6358 (0.68%)	**0.6613**
Seq-to-Seq					
MulT [126]	0.4590 (6.19%)	0.8022 (1.18%)	0.7951 (1.32%)	0.6980 (10.53%)	0.6511 (1.54%)
Ours					
QMF	0.4788 (2.15%)	0.8069 (0.60%)	0.7977 (0.99%)	0.6399 (1.33%)	0.6575 (0.57%)

Table 4.1: Effectiveness on CMU-MOSEI. The best scores out of all the models for a specific metric are in bold. The percentage difference from the best score (%Δ) is shown in parentheses next to the absolute performance of a model.

lead to a non-negligible drop on the 7-level accuracy. In fact, the 7-level accuracy on CMU-MOSI had the greatest *coefficient of variation* out of all metrics, suggesting low stability of this metric. On the larger dataset of CMU-MOSEI, QMF consistently outperformed MulT on all metrics, which was previously perceived as the best-performed model in this domain.

Model	Acc-7	Acc-2	F1	MAE	Corr
Vanilla LSTM					
EF-LSTM	0.3323 (9.90%)	0.7770 (2.90%)	0.7772 (2.60%)	0.9675 (5.78%)	0.6504 (6.54%)
LF-LSTM	0.3178 (13.83%)	0.7711 (3.63%)	0.7702 (3.48%)	0.9768 (6.80%)	0.6381 (8.31%)
LSTM+					
MARN [153]	0.3294 (10.68%)	0.7959 (0.54%)	0.7955 (0.31%)	0.9576 (4.70%)	0.6739 (3.16%)
MFN [152]	0.3236 (12.26%)	0.7851 (1.89%)	0.7838 (1.78%)	0.9684 (5.88%)	0.6380 (8.32%)
Tensor					
TFN [151]	0.3586 (2.77%)	0.7784 (2.72%)	0.7785 (2.44%)	0.9642 (5.42%)	0.6591 (5.29%)
LMF [85]	**0.3688**	0.7872 (1.62%)	0.7871 (1.37%)	0.9409 (2.88%)	0.6595 (5.23%)
Seq-to-Seq					
MulT [126]	0.3528 (4.34%)	**0.8002**	**0.7980**	0.9407 (2.86%)	0.6911 (0.69%)
Ours					
QMF	0.3353 (9.08%)	0.7974 (0.35%)	0.7962 (0.23%)	**0.9146**	**0.6959**

Table 4.2: Effectiveness on CMU-MOSI. The best scores out of all the models for a specific metric are in bold. The percentage difference from the best score (%Δ) is shown in parentheses next to the absolute performance of a model.

Another interesting finding is that different types of prior models, including advanced LSTM-based approaches, tensor-based models and seq2seq-based methods, were close to each other by effectiveness, and consistently outperformed the simple EF-LSTM and LF-LSTM strategies. Even

though similar trends have also been reported in the existing literature, the gaps observed in this experiment were much smaller. We conjecture that this is mainly because word embeddings were trained in this experiment while they were not set as trainable in previous models. Under a fixed word representation, the complexity of the neural structures may have an enormous impact on the representation capability of the model and hence influence performance. On the other hand, with trainable word embeddings, even a simple network structure may yield acceptable performance with a large number of training parameters in the embedding lookup table. This also explains why MulT was previously perceived as the best model but did not significantly outperform the remaining models in our experiment.

Empirically, we find that QMF is among the three best models, including MARN and LMF, which converges faster to better results at training when compared to other competitive approaches (see Figure 4.5). We illustrate the training time of QMF in minutes when compared to other SOTA on the CMU-MOSEI task in Figure 4.6. QMF attains a significant speedup during inference compared to more sophisticated models, such as MARM, MULT, and MFN, which are currently the SOTA approaches on multimodal sentiment analysis.

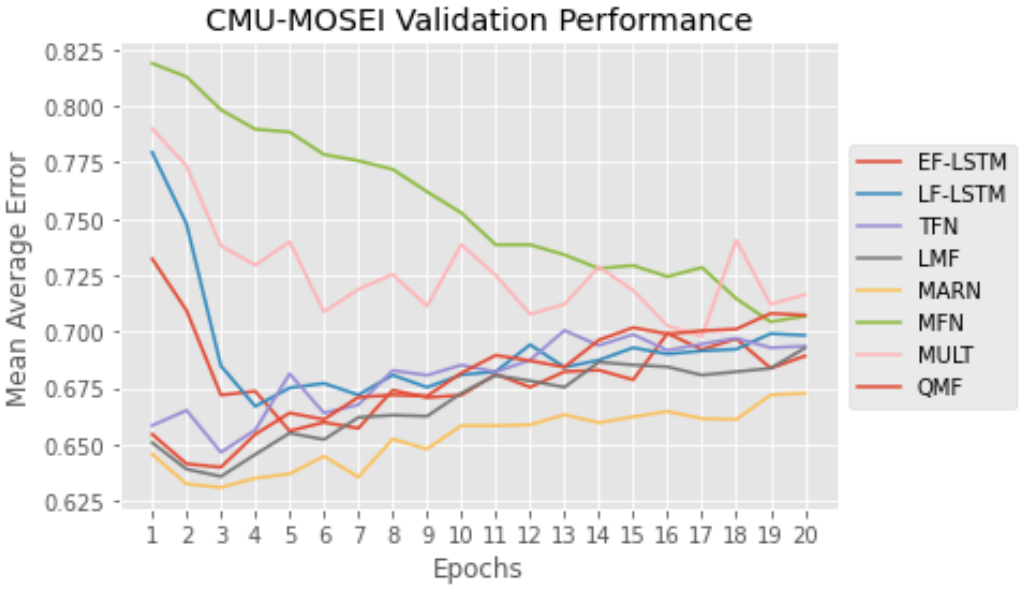


Figure 4.5: Validation set convergence of QMF in comparison with other SOTA on the CMU-MOSEI task.

4.4.3 Ablation Test

In order to examine the influence of each component in the proposed model, an ablation test was designed on the larger of the two datasets, i.e., CMU-MOSEI. Based on the best settings, changes

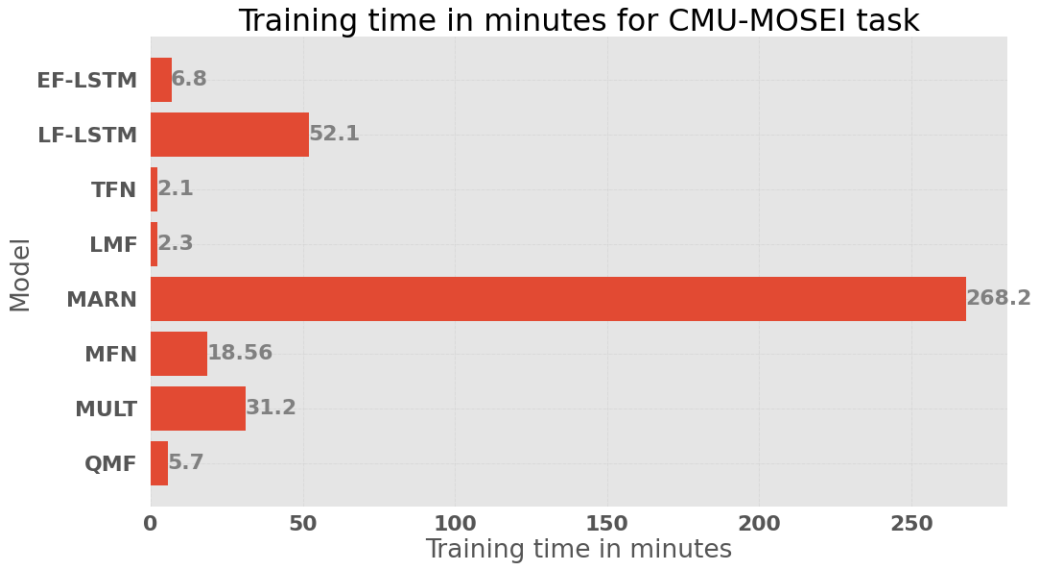


Figure 4.6: Training time of QMF in comparison with other SOTA on the CMU-MOSEI task.

were made only on the respective component, so that the performance difference was a reliable indicator of the impact of the element.

To validate the effectiveness of the modulus-argument assignment of complex-valued embedding, we replaced the complex-valued components with their real counterparts. However, simple removal of the arguments leads to a decrease in parameter scale and may bias the results. In order to eliminate this effect, the real-valued network QMF-real contained doubled dimensions $t_{dim} = 20, v_{dim} = 20, a_{dim} = 20$ for uni-modal inputs and twice the number of sentiment-related aspects $K = 40$.

A particular strategy was introduced to initialize the arguments of three modalities. To check whether it positively affected the model performance, we re-trained the same model with randomly initialized arguments (i.e., QMF-rand-init) and compared its performance with the original QMF model.

Another crucial network unit was the local-mixture strategy, where the density matrices of local contexts were extracted and fed to the measurement. To justify the use of this component, we run a model with a global mixture of all words in the sentence (i.e., QMF-global-mixture), with the other setting unchanged.

Finally, after the measurement results were outputted, a row-wise max-pooling was conducted to identify the most representative context for each sentiment-related aspect (i.e., eigenstate). We contrasted that strategy with the row-wise average-pooling (QMF-average-pool), which uses the average probability of all local contexts to represent the sentence feature with respect to a particular aspect.

As shown in Table 4.3, a notable drop in performance was observed for all QMF variants.

Models	Acc-7	Acc-2	F1	MAE	Corr
QMF	0.4788	0.8069	0.7977	0.6399	0.6575
QMF-real	0.4241	0.7301	0.7320	0.7641	0.4682
QMF-rand-init	0.4221	0.7172	0.7278	0.7583	0.5332
QMF-global-mixture	0.4324	0.7237	0.7244	0.7671	0.4215
QMF-average-pool	0.4208	0.7325	0.7401	0.7102	0.5542

Table 4.3: Ablation Study on CMU-MOSEI.

That illustrated the usefulness of complex-valued components, arguments initialization strategies, the local mixture strategy, as well as the max-pooling for measurement results. In particular, the discrepancy with QMF-real empirically implied that the complex values in the components were not merely a doubling of parameters. However, it brought about a meaningful combination of the respective features for the modulus and argument parts (which agrees with Section 4.3.1) that led to the performance improvement.

4.4.4 Interpretation of Multimodal Decision

The proposed model captures inter-modal interactions on the decision level, viewing the multimodal sentiment judgment as a composition of uni-modal decisions. In order to understand the *composited sentiment decision*, we decomposed the best QMF model on CMU-MOSEI by looking into the decisions on uni-modal and bi-modal data implicitly encoded in the tri-modal sentiment analysis. The decomposition approach was introduced in Section 4.3.5.

Models	Acc-7	Acc-2	F1	MAE	Corr
QMF-tri-modal	0.4788	0.8069	0.7977	0.6399	0.6575
QMF-textual	0.3644	0.7629	0.7064	1.1789	0.4949
QMF-visual	0.2893	0.4135	0.1299	0.9400	0.1608
QMF-acoustic	0.2897	0.4137	0.1301	0.9623	0.0313
QMF-textual+visual	0.3923	0.7973	0.7780	0.7800	0.5796
QMF-textual+acoustic	0.3955	0.7971	0.7748	0.7305	0.5509
QMF-visual+acoustic	0.2053	0.2897	0.1301	1.0731	0.2073

Table 4.4: Uni-modal and bi-modal sentiment classification result on CMU-MOSEI, entailed by the best-performed QMF learned by the whole CMU-MOSEI data.

Table 4.4 shows the sentiment prediction results based on all uni-modal and combinations of bi-modal features of CMU-MOSEI. The results showed that the QMF best predicted the sentiment based on three modalities. When QMF was used to predict sentiment based on uni-modal data, it was able to give a reasonably accurate prediction for textual features, but barely able to provide any predictions for visual and acoustic modalities. However, the QMF was able to give better judgments when combining visual or acoustic features with textual features, as can be seen from the gradually

increased performances in both textual → textual+visual → textual+visual+acoustic and textual → textual+acoustic → textual+visual+acoustic paths.

The results above indicate that the textual modality played a predominant role in determining the sentiment, while visual and acoustic modalities were less relevant to the sentence sentiment. This finding is consistent with the prior work in this field [106, 126, 151]. Furthermore, even if the textual modality carried the majority of sentiment-related information, complementary information was extracted from the visual and acoustic modalities to boost the sentiment prediction capability of textual modality. However, visual and acoustic modalities were not able to predict sentiment independently.

Table 4.5 shows examples from the CMU-MOSI dataset to show the impact of our proposed QMF multimodal fusion approach by comparing with uni-modals and bi-modals. Each example is described with the spoken words as well as the acoustic and visual behaviours. The sentiment predictions and the ground truth labels range between strongly negative (-3) and strongly positive (+3).

Table 4.6 shows some utterances illustrating how local word contexts, captured by sliding windows of varying lengths through the whole sentence, i.e., n-grams, contribute to the final sentiment analysis. Since the compact multimodal representation in the entire multimodal space is tough to understand for humans, we illustrate cases for the textual modality only by decomposing the multimodal representation to the textual Hilbert space. The ground truth sentiment labels in Table 4.6 are between strongly negative (-3) and strongly positive ($+3$). To this goal, we filter local word contexts by filtering out probability scores from the row-wise max-pooling layer (see Figure 4.4), which are less than 0.2. The darker the colour in Table 4.6, the bigger weight of a sliding window is. Overall, we noticed that context windows with a longer length, e.g., trigrams, contribute most to the final sentiment polarity of the utterances. However, in some cases, the strong polarity of some words dominates the local context of more expansive sliding windows. For example, in the third utterance (see Table 4.6), the word "good" dominates the local window context "the good evil kind". We also notice that some textual contexts which are neutral themselves, e.g., "Israel sniper" in the last utterance (see Table 4.6), convey an increased sentiment polarity. For those cases, we attribute the polarity to the interaction with other modalities.

Spoken words + acoustic and visual behaviors	Textual	Visual	Acoustic	Textual, Visual	Textual, Acoustic	Visual, Acoustic	QMF	Ground Truth
It is very very loyal to the book Excited voice	1.7	-0.4	2.4	1.4	2.6	1.2	1.5	1.6
I just had fun watching the movie Smile expression Loud voice	0.6	1.3	1.4	1.3	1.6	1.9	2.0	2.2
I do not understand this movie Frowning expression	-1.1	-1.5	0.4	-1.7	-1.4	-1.1	-1.9	2.2
The only actor who can really sell their lines is Erin Eckart Frown Low-energy voice	0.1	-0.6	-0.7	-0.6	-0.7	-0.8	-0.9	-1.0

Table 4.5: Examples from the CMU-MOSI dataset. The ground truth sentiment labels are between strongly negative (-3) and strongly positive (+3). For each example, we show the prediction output of uni-modals, bi-modals, and tri-modals.

Utterance	Ground Truth
So it is perfect for kids as well as just like teenagers as well who just wanna see a lot of explosions and some crazy stuff.	2.2
I just got finished watching an excellent movie called Mars Needs Moms.	3.0
There is not in my opinion there wasn't very much of a balance between sort of like the good evil kind of thing.	-1.6
But just aside from that they really dropped the ball on the suspenseful parts.	-2.2
I liked the girl who was the Israeli sniper.	0.8

Table 4.6: Contribution of varying length sliding windows to the final sentiment analysis of utterances for CMU-MOSI task. The darker the colour, the bigger weight of a sliding window is. The labels are between strongly negative (-3) and strongly positive (+3).

Table 4.7 illustrates cases for the CMU-MOSI task, showing how the sentiment polarity of words, encoded into phases, changes after model learning. The "Initialization" column (Table 4.7) shows the prior polarity of words for each modality, as described in Section 4.3.1. The "Learning" column (Table 4.7) shows how the phase, and hence the sentiment polarity of specific multimodal words, changes after model learning. To extract the learning phases, we obtained unigrams of particular words after the interaction of multimodal representation with the learnable abstract sentiment concepts, i.e., quantum measurements (see Figure 4.4). The case studies in Table (Table 4.7) show that the phase learning process mitigates the overall polarity of words. For example, the sentiment polarity of words "surprised" and "ridiculous", initialized as strongly positive (i.e., 0) and strongly negative (i.e., π), changes to a moderate sentiment polarity after model learning, i.e., $\frac{\pi}{6}$ and $\frac{5\pi}{6}$ correspondingly, which is closer to the ground truth sentiment. In contrast, when the sentiment polarity of a word is neutral, e.g., the word "sell" in the last case (Table 4.7), the phase changes its sentiment polarity accordingly, after the interaction of modalities and during the learning process.

	Phases		
Spoken words + visual behaviours and acoustic	**Initialization**	**Learning**	**Ground Truth**
[It actually surprised me.] + Neutral Voice + Neutral Expression	{0,0,0}	$\approx \frac{\pi}{6}$	1.2
[The plot is ridiculous but no one took it seriously] + Neutral Voice + Neutral Expression	$\{\pi,\pi,\pi\}$	$\approx \frac{5\pi}{6}$	-1.4
[The only actor who can really sell their lines is Erin Eckart.] + Frown + Low-energy voice	$\{\frac{\pi}{2},\frac{\pi}{2},\frac{\pi}{2}\}$	$\approx \frac{2\pi}{3}$	-1.0

Table 4.7: Cases from CMU-MOSI task. For each case, we show how the initialization of uni-modal phases, defined in {}, for the corresponding marked textual words (e.g., *surprised*, *ridiculous*, and *sell*), change after model learning. The phases in the third column correspond to the unified phases of the compact multimodal representation. The ground truth sentiment labels are between strongly negative (-3) and strongly positive (+3).

4.5 Conclusions

We have developed a novel quantum-inspired framework for multimodal sentiment analysis. The framework borrows quantum concepts to explicitly model intra-modal interactions on the feature level and inter-modal interactions on the decision level via the concept of superposed and composite states, respectively. A neural network with complex-valued components was built to learn both interactions in an end-to-end supervised way. In addition to obtaining comparable performance to

SOTA models, the model facilitated to understanding of multimodal interactions from both quantum and classical perspectives.

Although our results are encouraging, the model is subject to multiple limitations. First, the model captures inter-modal interactions among different modalities leaving aside speaker information and relative position of preceding utterances, i.e., a.k.a., context modelling. Crucially, the model considers word-level interactions via the concepts of mixture and superposition on the feature level, while the interactions across different modalities are largely absent from the feature level. The multimodal representation is hence a separable state rather an entangled representation with respect to the three uni-modal modalities. This means that the current strategy has not fully exploited the expressive power of quantum probabilities to fuse inputs of multimodal features. In the next chapter, we take a step forward and present a comprehensive investigation to encode cross-modal information in the form of non-classical correlations, i.e., a.k.a., *entanglement*.

Chapter 5

Entanglement-driven Multimodal Fusion

A multimodal fusion process requires the consideration of distinctive issues. Among them is the correlation among different modalities representing how they co-vary with each other. In many situations, the correlation between them provides additional cues that are very useful in fusing them. In particular, the basic idea of correlation is that a modality can help predict or enhance another modality. It is hence essential to know different methods of computing correlations and to analyze them from the perspective of how they affect fusion [110]. The correlation can be comprehended at various levels, e.g. the correlation between low-level features and the correlation between semantic-level decisions. The correlation between features has been computed in the forms of correlation coefficient, mutual information, latent semantic analysis (also called lament semantic indexing), canonical correlation analysis, and cross-modal factor analysis [11]. On the other hand, the decision level correlation has been exploited in the form of causal link analysis, causal strength, and agreement coefficient [11].

In quantum mechanics, correlation has also been an important topic. In the quantum mechanical framework, uncertainty may occur not only when the elements are collected in an ensemble but also when each of them is in a superposed state. In QT, making an observation on one part of a system *instantaneously* could affect the state in another part of a system, even if space-like distances separate the respective systems. Such a quantum correlation presents some peculiarities, which led to the notion of *entanglement*. Entanglement is a sort of correlation between observables measured in atomic-size particles, such as photons, when these particles are not necessarily collected in ensembles.

Despite entanglement being a kind of correlation, there are some fundamental differences between entanglement and the classical correlation encountered in the macroscopic world. A classical correlation is a statistical relationship, causal or not, between two random variables. In entanglement,

besides correlation, cause exists as well since the correlation does not depend on an underlying value attached to the particles. Instead, it depends on what is measured on either side. This non-classical property of quantum entanglement motivates us to model non-classical correlations in the multimodal fusion process and investigate causal relationships between different modalities, facilitating hence explainability of learning models.

Despite the recent advances in quantum probabilistic neural models, in particular the quantum-inspired multimodal fusion (QMF) framework presented in Chapter 4, existing models treat quantum states as either a classical mixture or as a decomposable tensor product across modalities, without triggering their interactions which could make them correlated or non-separable (i.e., entangled). This means that the current strategies have not fully exploited the expressive power of quantum probabilities. Such non-separability has been shown in cognitive science as a fundamental feature of human decision making under uncertainty, and it complies with the more general quantum probability theory only. To fill this gap, in this chapter, we investigate the encoding of cross-modal information in the form of non-classical correlations, a.k.a., entanglement.

In this chapter (Chapter 5), we first present a prior study to examine a multimodal fusion scenario that might be similar to that encountered in physics by firstly measuring two observables of a multimodal document, i.e., text-based and image-based, without counting on an ensemble of multimodal documents already labelled in terms of these two variables. Then, we investigate the existence of non-classical correlations between pairs of uni-modal decisions. The experimental results and discussions provide theoretical and empirical insights and inspirations to develop a transparent and joint entanglement-driven fusion neural network for emotion recognition in conversations, addressing not only the challenges of multimodal fusion but also those of context modelling in the form of preceding utterances. The last part of chapter 5 presents an evaluation of the proposed entangled-driven fusion neural network on the video sentiment analysis task. Overall, the model achieves a performance improvement and optimized post-hoc interpretability via the notion of entanglement for both video sentiment analysis and emotion recognition tasks.

5.1 Investigating Non-classical Correlations Between Decision Fused Multimodal Documents

5.1.1 Introduction

Nowadays, images and text are an integral part of the Web, where images rarely exist without text. However, Web search systems still consider images as a separate vertical from text and provide only text-to-image search functionality. Yet, the spectrum of information needs of Web search system users goes well beyond text-to-image searches and includes the search tasks, in which pairs of a textual fragment and an image form *atomic retrieval units*, such as the text-to-image and text IR

scenario. For instance, suppose a user types in a text query to retrieve multimodal documents

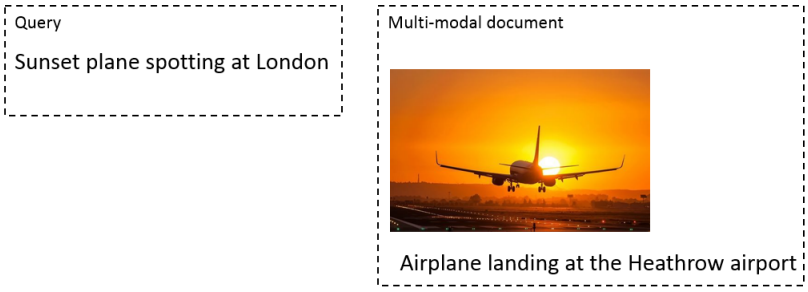


Figure 5.1: The text-to-image and text IR scenario

consisting of an image and a caption as shown in Fig. 5.1. One can notice that the query term "plane" can be matched in both textual and visual modalities of the given multimodal document. However, the query term "London" can be matched only in its textual modality, while the term "sunset" only in its visual modality. This implies that only when the text and image modalities are fused, we get the benefit of complementary information, in turn increasing the precision of IR. It is hence natural to consider both text and image as one retrieval unit. To this end, feature-level or decision-level fusion approaches (Section 1.1) should be in place.

In QT, an entangled composite system cannot be validly decomposed and modelled as separate sub-systems (Section 3.2.7). If a composite system is factorizable, then it is compositional, in a sense it can be expressed as a product of states corresponding to the separate sub-systems. A system that is not factorizable cannot be described by its individual sub-systems, and is deemed non-compositional, termed *entangled* [99]. That is, individual sub-systems are viewed as one unit. QT provides a well-developed set of analytical tools that can be used to determine whether the state of a system of interest can be validly decomposed into separate sub-systems. A possible way to test the non-compositional state of a composite system is the violation of Bell's inequalities. For instance, having calculated the expectation values of variables associated with an experiment, we can fit the Clauser-Horne-Shimony-Holt (CHSH) version of Bell's inequality [35]. If the CHSH inequality is greater than 2, then the Bell inequality is violated. It has been empirically found that the maximal possible violation in QT is $2\sqrt{2} \approx 2.8284$ [34]. This means that each violation being close to the maximal value is very significant. In addition to the CHSH inequality, the Schmidt decomposition is another way for detecting entanglement in bipartite systems (Section 3.2.7).

In the current preliminary study, we have modelled image and text relevances of documents concerning multimodal queries as a composite system, and then investigated the existence of non-classical correlations between image and text relevance probabilities via the violation of CHSH version of Bell's inequality. The contribution of this work resides more in a theoretical than a practical aspect. In particular, Section 5.1 presents a decision-level modality fusion strategy with a quantum-inspiration, while the rather challenging feature-level fusion is addressed in Section 5.2

and 5.3.

5.1.2 Background

We now briefly introduce the fundamental concepts of quantum entanglement, which have been exploited to model image and text decisions and investigate the existence of entanglement between pairs of documents.

Quantum States

At the outset, let us suppose a system of two qubits expressed in a Bit basis $\{0, 1\}$, such that the first qubit is in a state $|\psi_A\rangle = a_0 |0\rangle + a_1 |1\rangle$ and the second one in a state $|\psi_B\rangle = b_0 |0\rangle + b_1 |1\rangle$. The state $|\psi\rangle$ of the two qubits together as a composite system is a superposition of four classical probabilities resulting in

$$|\psi\rangle = |\psi_A\rangle \otimes |\psi_B\rangle = a_0 b_0 |00\rangle + a_0 b_1 |01\rangle + a_1 b_0 |10\rangle + a_1 b_1 |11\rangle , \tag{5.1}$$

where $\otimes$ denotes the tensor product of states.

Let now assume that the composite quantum state $|\psi\rangle$ is in an entangled state, given by the following Bell state

$$|\hat{\psi}\rangle = \frac{1}{\sqrt{2}} |00\rangle + \frac{1}{\sqrt{2}} |11\rangle . \tag{5.2}$$

According to Born's rule (Section 3.2.4), the probability of the composite quantum state $|\hat{\psi}\rangle$ to collapse either to the state $|00\rangle$ or to the state $|11\rangle$ equals 0.5. However, after the measurement, the composite system is not in an entangled state anymore. For instance, once we measure the state $|00\rangle$, the updated state of the system results in

$$|\hat{\hat{\psi}}\rangle = |00\rangle . \tag{5.3}$$

Moreover, if we first measure the state $|0\rangle$ of the first qubit (Equation 5.2), the probability of the qubit to collapse to the state $|0\rangle$ again equals 0.5. However, after the measurement, the probability of the second qubit to be in the state $|0\rangle$ results in 1.

Let us consider the scenario of changing bases from the Bit basis $\{0, 1\}$ to a Sign basis $\{-, +\}$. According to the rotation invariance [123], the Bell state in the Sign basis is again an equal superposition of the state $|--\rangle$ and the state $|++\rangle$ such that

$$\begin{aligned} |\hat{\psi}\rangle &= \frac{1}{\sqrt{2}} |00\rangle + \frac{1}{\sqrt{2}} |11\rangle \\ &= \frac{1}{\sqrt{2}} |--\rangle + \frac{1}{\sqrt{2}} |++\rangle . \end{aligned} \tag{5.4}$$

Suppose now that we measure the probability of the second qubit to be in the state $|-\rangle$ in respect of the Sign basis, given that we have already measured the probability of the first qubit to be in state $|0\rangle$ concerning the Bit basis. Once we measure the first qubit, the probability of the second qubit to be in the same state $|0\rangle$ results in 1. If θ is the angle between the Bit and Sign bases, then according to the Pythagorean theorem, the probability of the second qubit to be in the state $|-\rangle$ equals $\cos^2\theta$.

Bell Inequality

In quantum mechanics, one way to test entanglement is by violating Bell's inequalities. A possible strategy to proceed is to define four observables. Each observable has binary values ± 1 thus gives two mutually exclusive outcomes. For instance, a photon can be detected by "+" or "-" channel in a two-channel polariser (see Figure 5.2). If A_1, A_2 are observables describing the first system, and B_1, B_2 observables of the second system, respectively, then the CHSH inequality has as follows:

$$|\langle A_1 B_1\rangle + \langle B_1 A_2\rangle + \langle A_1 B_2\rangle - \langle A_2 B_2\rangle| \leq 2, \tag{5.5}$$

where $\langle\rangle$ denotes the expectation value[1] between two observables. The violation of 5.5 is a sign of entanglement. A Bell inequality violation implies that at least one of the assumptions of *local-realism* made in the proof of 5.5 must be incorrect [99]. This points to the conclusion that either or both of locality - an object is only directly influenced by its immediate surroundings - and realism - an object has definite values - must be rejected as a property of composite systems that violate CHSH inequality.

5.1.3 Model

The experimental setup is analogous to that one of investigating quantum entanglement in photons [10]. Figure 5.2 shows a typical optical experiment of the two-channel Bell test. In particular,

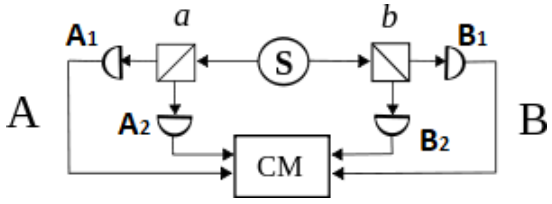


Figure 5.2: A typical CHSH (two-channel) experiment. The source S produces pairs of "photons", sent in opposite directions. Each photon encounters a two-channel polariser whose orientation can be set by the experimenter. Emerging signals from each channel are detected and coincidences counted by the coincidence monitor CM.

[1]More details about the calculation of expectation values are given in Section 5.1.3

documents corresponds to photons, while the relevance and non-relevance of documents are analogues to the mutually orthogonal outcomes of the two-channel polariser. Finally, the functionality of the monitor CM (see Figure 5.2) in the lab experiment is similar to the decision-level modality fusion process.

Preparation of Quantum States

The proposed method draws an analogy from quantum composite systems to model image and text modality decisions of relevance in respect of a multimodal query in a multimodal IR scenario. Before the decision-level fusion of image and text modalities, there exists a probability $P(R|T)$ for a multimodal document D_M to be relevant to a multimodal information need concerning the textual information. Similarly, the probability for the same document not to be relevant is denoted as $P(\overline{R}|T)$, which is equal to $1 - P(R|T)$. Let us consider a real-valued two dimensional Hilbert space for the relevance of the D_M concerning the textual information, as illustrated in Figure 5.3. The vector R_t stands for the relevance of the document concerning the text-based modality. On the other hand, $\overline{R_t}$ represents the non-relevance concerning the same text-based information need and is orthogonal to R_t.

The text-based relevance of a document can be modelled as a vector in the Hilbert space, which unifies logic, probabilities, and geometry into a compact vector space [134]. In particular, the vector is a superposition of the relevance and non-relevance basis vectors with respect to the text-based modality and is represented as:

$$|D_M\rangle = a\,|R_t\rangle + a'\,|\overline{R_t}\rangle\,, \tag{5.6}$$

where $|a|^2 + |a'|^2 = 1$. The coefficients a and a' are calculated by projecting the quantum state $|D_M\rangle$ onto the relevance and non-relevance basis vectors respectively (Figure 5.3). Mathematically speaking, a equals the square of the inner product of vectors $|D_M\rangle$ and $|R_t\rangle$. Hence, according to the Born rule (Sec 3.2.4), $P(R|T)$ equals the square of the inner product of vectors $|R_t\rangle$ and $|D_M\rangle$, i.e., $|\langle R_t|D_M\rangle|^2$. Likewise, $P(\overline{R}|T)$ equals $|\langle\overline{R_t}|D_M\rangle|^2$.

Similarly, we denote as $P(R|I)$ the probability of the multimodal document D_M to be relevant concerning the image-based information need, and $P(\overline{R}|I)$ the probability to be irrelevant respectively (Figure 5.4). The relevance of a document with respect to the image-based modality is similarly modelled as:

$$|D_M\rangle = b\,|R_i\rangle + b'\,|\overline{R_i}\rangle \tag{5.7}$$

Thus, $P(R|I)$ is computed as the square of the inner product of vectors $|R_i\rangle$ and $|D_M\rangle$, i.e., $|\langle R_i|D_M]\rangle^2$. Likewise, $P(\overline{R}|I)$ equals $|\langle\overline{R_i}|D_M\rangle|^2$.

By contrast, after the decision-level fusion of image and text modalities, the document is judged based on both modalities. Such a phenomenon can be modelled in the same Hilbert space, spanned by two different bases, one for each modality, as illustrated in Figure 5.5. The document D_M is

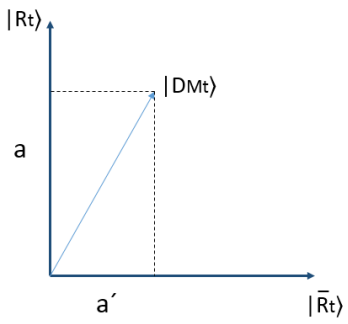


Figure 5.3: Hilbert space of text-based relevance representation.

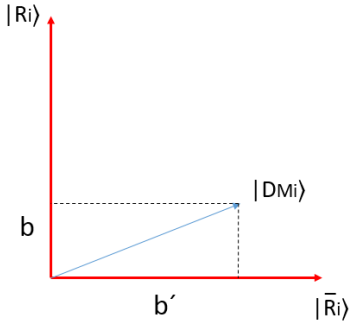


Figure 5.4: Hilbert space of image-based relevance representation.

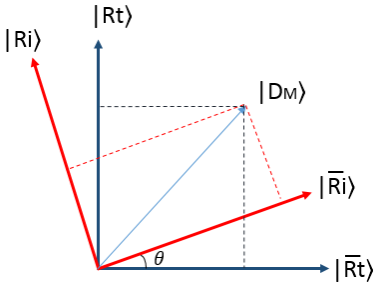


Figure 5.5: Hilbert space of multimodal relevance representation.

hence represented as a unit vector and its representation is expressed with respect to the bases $T = \{|R_t\rangle, |\overline{R_t}\rangle\}$ and $I = \{|R_i\rangle, |\overline{R_i}\rangle\}$. Note that each basis models context with respect to a given modality.

Quantum-inspired Decision-Level Fusion

After the preparation of quantum states, we have modelled the decision-level fusion of pairwise documents as a quantum composite state. For a pair of documents, namely, D_{M_1} and D_{M_2}, where each document is in a superposition state of relevance and non-relevance, spanned by the text and image bases, the composite state is defined by the tensor product of the document quantum states. Mathematically speaking, if $|D_{M_1}\rangle = a_1 |R_t\rangle + a_1' |\overline{R_t}\rangle$ and $|D_{M_2}\rangle = a_2 |R_t\rangle + a_2' |\overline{R_t}\rangle$ are the quantum states of documents D_{M_1}, D_{M_2}, in respect to the text modality, then the composite state $|\hat{D_M}\rangle$ has as follows:

$$|\hat{D_M}\rangle = |D_{M_1}\rangle \otimes |D_{M_2}\rangle = a_1 a_2 |R_t R_t\rangle + a_1 a_2' |R_t \overline{R_t}\rangle + a_1' a_2 |\overline{R_t} R_t\rangle + a_1' a_2' |\overline{R_t R_t}\rangle , \tag{5.8}$$

where $\otimes$ the tensor product of quantum statets. Similarly, if $|D_{M_1}\rangle = b_1 |R_i\rangle + b_1' |\overline{R_i}\rangle$ and $|D_{M_2}\rangle = b_2 |R_i\rangle + b_2' |\overline{R_i}\rangle$ are the quantum states of documents D_{M_1}, D_{M_2}, in respect to the image modality, then the composite state $|\hat{D_M}\rangle$ has as follows:

$$|\hat{D_M}\rangle = |D_{M_1}\rangle \otimes |D_{M_2}\rangle = b_1 b_2 |R_i R_i\rangle + b_1 b_2' |R_i \overline{R_i}\rangle + b_1' b_2 |\overline{R_i} R_i\rangle + b_1' b_2' |\overline{R_i R_i}\rangle . \tag{5.9}$$

If the composite state of documents $|\hat{D_M}\rangle$ is factorizable, there is an uncertainty concerning the relations between the two documents. For instance, the first and second terms in Equation 5.8 reveal that when the linguistic content of the D_{M1} is relevant, then we cannot be sure about the relevance of the linguistic content of the other document since it could be relevant or non-relevant. Similarly, the third and fourth term in Equation 5.8 implies that when the linguistic content of the D_{M1} is

non-relevant, then the other document is in a superposition of relevance and non-relevance basis states. Likewise, image modality raises similar ambiguities. Moreover, when the composite state of documents is factorizable then it is compositional in a sense it can be expressed as a product of individual states, corresponding to the separate documents D_{M_1}, D_{M_2}.

Conversely, if the composite quantum state is not factorizable is deemed *non-compositional* and termed *entangled*. In this case, if the textual basis is a the standard basis, then for each pair of documents, there exists one of the two following Bell states:

$$|D_M\rangle = a_1 a_2 \, |R_t R_t\rangle + a'_1 a'_2 \, |\overline{R_t R_t}\rangle \, , \tag{5.10}$$

or

$$|D_M\rangle = a_1 a'_2 \, |R_t \overline{R_t}\rangle + a'_1 a_2 \, |\overline{R_t} R_t\rangle \, . \tag{5.11}$$

If the first Bell state exists (Equation 5.10), then the probability for both documents to be relevant regarding the linguistic modality equals $|a_1 a_2|^2$. If we measure only the probability of the first document to be relevant in respect to the linguistic modality, it results again in $|a_1 a_2|^2$. Then after the measurement, the probability for the second document to be relevant is equal to 1. Moreover, we can *simultaneously* predict the probability of the second document to be relevant concerning the image modality, which is equal to $cos^2\theta$, where θ is the angle between the image and text bases (see Figure 5.5). Likewise, we can similarly estimate the probabilities of both documents to be irrelevant (e.g., $|\overline{R_t R_t}\rangle$, Equation 5.10), the first to be relevant and the second irrelevant (e.g., $|R_t \overline{R_t}\rangle$, Equation 5.11), or the first to be irrelevant and the second relevant (e.g., $|\overline{R_t} R_t\rangle$, Equation 5.11).

CHSH Inequality Estimation

The CHSH inequality defines four observables, where each observable has two binary values ± 1 thus gives two mutually exclusive outcomes. In particular, for the document D_{M1} there exist two variables, R_{t_1} and R_{i_1}, which take the values $1, -1$ respectively, where $R_{t_1} = 1$ corresponds to the basis state $|R_{t_1}\rangle$ and $R_{t_1} = -1$ corresponds to its orthogonal basis state $|\overline{R_{t_1}}\rangle$. Similarly, $R_{i_1} = 1$ corresponds to the basis state $|R_{i_1}\rangle$ and $R_{i_1} = -1$ corresponds to its orthogonal basis state $|\overline{R_{i1}}\rangle$. For the document D_{M2}, we have variables R_{t_2} and R_{i_2} which also take values $1, -1$, where $R_{t_2} = 1$ corresponds to the basis state $|R_{t_2}\rangle$ and $R_{t_2} = -1$ corresponds to its orthogonal basis state $|\overline{R_{t2}}\rangle$. Similarly, $R_{i_2} = 1$ corresponds to the basis state $|R_{i2}\rangle$ and $R_{i2} = -1$ corresponds to its orthogonal basis state $|\overline{R_{i_2}}\rangle$. Then, according to Equation 5.5, the CHSH inequality has as follows:

$$|\langle R_{t_1} R_{t_2}\rangle + \langle R_{t_2} R_{i_1}\rangle + \langle R_{t_1} R_{i_2}\rangle - \langle R_{i_1} R_{i_2}\rangle| \leq 2, \tag{5.12}$$

where $\langle\rangle$ denotes expectation values of observable, which are calculated[2] as follows:

$$\begin{aligned}\langle R_{t1}R_{t2}\rangle &= ((+1)P(R_{t1}) + (-1)P(\overline{R_{t1}})) * ((+1)P(R_{t2}) + (-1)P(\overline{R_{t2}})) \\ &= P(R_{t1})P(R_{t2}) - P(R_{t1})P(\overline{R_{t2}}) - P(\overline{R_{t1}})P(R_{t2}) + P(\overline{R_{t1}})P(\overline{R_{t2}}),\end{aligned}$$

$$\begin{aligned}\langle R_{t2}R_{i1}\rangle &= ((+1)P(R_{t2}) + (-1)P(\overline{R_{t2}})) * ((+1)P(R_{i1}) + (-1)P(\overline{R_{i1}})) \\ &= P(R_{t2})P(R_{i1}) - P(R_{t2})P(\overline{R_{i1}}) - P(\overline{R_{t2}})P(R_{i1}) + P(\overline{R_{t2}})P(\overline{R_{i1}}),\end{aligned}$$

$$\begin{aligned}\langle R_{t1}R_{i2}\rangle &= ((+1)P(R_{t1}) + (-1)P(\overline{R_{t1}})) * ((+1)P(R_{i2}) + (-1)P(\overline{R_{i2}})) \\ &= P(R_{t1})P(R_{i2}) - P(R_{t1})P(\overline{R_{i2}}) - P(\overline{R_{t1}})P(R_{i2}) + P(\overline{R_{t1}})P(\overline{R_{i2}}),\end{aligned}$$

$$\begin{aligned}\langle R_{i1}R_{i2}\rangle &= ((+1)P(R_{i1}) + (-1)P(\overline{R_{i1}})) * ((+1)P(R_{i2}) + (-1)P(\overline{R_{i2}})) \\ &= P(R_{i1})P(R_{i2}) - P(R_{i1})P(\overline{R_{i2}}) - P(\overline{R_{i1}})P(R_{i2}) + P(\overline{R_{i1}})P(\overline{R_{i2}}),\end{aligned}$$

where $P(\cdot)$ stands for probability of an outcome. Note that the products of probabilities are defined as joint probabilities between two independent outcomes. Moreover, the violation of Equation 5.12 is a sign of entanglement, and then a pair of documents may result in one of the Bell states of Equations 5.10 or 5.11.

5.1.4 Methodology

Dataset

We have conducted experiments on the ImageCLEF2007 data collection [56], the purpose of which is to investigate the effectiveness of combining image and text for IR tasks. Out of 60 test queries, we randomly picked up 30 ones, together with the ground truth data. Each query, describing user information need, consists of three sample images and a text description, whereas each document consists of an image and a text description. For every query, we created a subset of 300 relevant and irrelevant documents, which firstly includes all the relevant documents for the query, and the rest is irrelevant documents. The dataset has been used for investigating both the Bell states (Equations 5.10 and 5.10). The number of relevant documents per query ranges from 11 to 98.

Estimation of Uni-modal Retrieval Scores

For image modality, we exploited the pre-trained VGG16 model [121] on ImageNet to extract features from images. The output is a 2049-dimensional vector representation. After the feature extraction

[2]For derivation look at Appendix B.

process, we computed similarity scores between a query and an image via cosine similarity. For the linguistic modality, a query expansion approach has been applied, extending a query with the ten most frequent terms according to the ground truth documents. This indeed corresponds to a simulated explicit relevance feedback scenario. Then, we mapped text to TF-IDF vector representations and compute similarity scores again via cosine function. Note that cosine similarity score is bounded in $[0, 1]$. However, in this study, cosine similarity is used to approximate the probability of relevance.

Experimental Procedure

At the outset, image and text relevance scores are estimated via cosine function. Then, expectation values are computed according to relevance scores. In particular, the probability of a document to be relevant concerning a modality is equal to the outcome of cosine function. Consequently, the likelihood of a document to be irrelevant concerning the same modality equals 1 minus the score of cosine function. Next, we fit the CHSH inequality (Equation 5.12) with the estimated expectation values and checked for any existence of violation. For each query, we calculated in total the percentage of documents that violated the CHSH inequality. Additionally, we calculated the rate of queries that violated the CHSH inequality.

5.1.5 Results and Discussion

The experiment results showed that there is no violation of Bell's inequality. This implies that in the context of the experimental setting non-classical correlations between pairs of documents may not exist, but also that the hypothesis of rotation invariance falls down. Thus, the image and text bases are not equal Bell states as defined in Equation 5.4.

Another possible explanation of the experimental results is that the outcomes of the observables are initially independent. For instance, the probability of the linguistic relevance of the first document does not affect the linguistic relevance probability of the second document. That is, the joint probability of relevance has been calculated as a product of individual relevance probabilities. By contrast, in some user studies the Bell inequality has been violated [5, 6, 21, 22]. In those studies, users are commonly asked to report their judgments about intersections of events. Hence the joint probabilities can be directly estimated from the judgments. Thus, the expectation values are calculated under an implicit assumption that the outcomes can be incompatible. This assumption may result in "conjunction fallacy" [129] violating the monotonicity law of probability by overestimating the joint probability, thus violating the Bell inequality.

Finally, another important reason is that we constructed real-value Hilbert spaces to represent the quantum states. However, the complex field is pivotal in QT because the quantum operators create superposition that can only be modelled in the complex field [1, 94].

5.1.6 Section Conclusions

We introduced a quantum-inspired methodology to investigate the existence of non-classical correlations between pairs of decision fused multimodal documents. The practical significance of such correlations is beyond the classical correlations. That is, if we know the relevance of an entangled document, then we can instantaneously know with certainty the relevance of the other entangled document, acquiring information about how to fuse local decisions. However, the experiments showed no violation of the CHSH inequality. We attribute this result to the lack of interactions between of documents after the integration of them in a composite system and the lack of complex field.

Overall, the contribution of this work resides more in a theoretical than a practical aspect. The experimental results and discussions provide theoretical and empirical insights and inspirations for future development of this direction. In particular, in the next section, we introduce a transparent and end-to-end complex-valued neural model for emotion recognition in conversations. The model induces different modalities to interact in such a way that they may not be separable, encoding cross-modal information in the form of non-classical correlations. Moreover, it models interactions across tri-modals, which to our best knowledge, no existing models in the current literature have taken into account.

5.2 An Entanglement-driven Neural Network Model for Contextual and Non-Separable Modality Fusion in Conversational Emotion Recognition

In chapter 4, we have developed a novel quantum-inspired framework for multimodal sentiment analysis. The framework considers word-level interactions via the concepts of mixture and superposition on the feature level, while the interactions across different modalities are largely absent from the feature level. Moreover, the model treats cross-modal interactions as a separable tensor product of modalities and interactions of words in a sentence as a classical mixture, which cannot fully exploit the potential of a quantum probabilistic description. Finally, the model does not consider contextual information in the form of preceding utterances. In Set. 5.1, we have presented a preliminary study to investigate non-classical correlations between decision fused multimodal documents. However, due to the lack of interactions between pairs of documents and complex field, non-classical correlations are absent. A full quantum model is yet to be developed to capture the non-classical correlations across distinct modalities explicitly. Such non-classical correlations are crucial for multimodal representation learning tasks, since one modality would influence the other in the process of reaching a final decision in a way that complies with quantum probability theory.

In this section, we develop a transparent and joint quantum probabilistic neural model, namely,

contextual Entanglement-driven Fusion Neural Network (c-EFNN). c-EFNN induces different modalities to interact in such a way that they may not be separable, encoding cross-modal information in the form of non-classical correlations. Evaluation on two benchmarking datasets for video emotion recognition in conversations shows improved performance over a wide range of SOTA baselines. Additionally, the degree of non-separability between modalities optimizes the post-hoc interpretability. The model is fundamentally different from that one in chapter 4 in that we take a quantum-cognitively motivated view on the non-decomposability of cross-modality interactions, which is modelled as quantum entanglement, while the model in chapter 4 assumed the interactions are decomposable.

5.2.1 Introduction

Understanding human-like emotions requires the consideration of behavioural cues (e.g., vocal behaviour, posture, gaze etc.), contextual information, and cognitive biases, such as moods and types of behaviour, mental states, or events influencing emotions [102, 116]. In recent years, research has made significant strides towards the inference, recognition, and interpretation of human-like emotions. In particular, neural approaches have been investigated to either model interactions across distinct modalities, i.e., linguistic, visual and acoustic, [126, 106, 143, 152, 151], or model interactions across parties in a conversation [59, 60, 92, 48], after merging different modalities into a joint multimodal representation. Although such approaches have demonstrated excellent performance, they neglected how people understand and reason about emotional states [102]. There is hence a need to distil cognitive biases into workable computational models.

As discussed in Section 1.1 (Multimodal Intelligence), the modelling of distinct modalities for emotion recognition is a challenging problem, due to the spectrum of emotions that an utterance can emerge, depending on the context of individual modalities. This implies that modality fusion is contextual and distinct modalities should not be considered in isolation; rather, each modality acts as a context for the other modalities. That is, we must model modalities as *non-separable*. QT is the only theory which models non-separability. Thus, there is a reason to suppose that QT provides an adequate theory to capture cross-modal correlations and how such correlations influence an emotional state.

QT is not only a physical theory but also a framework in which theories can be developed. Indeed, QT has increasingly been deployed outside physics. Early work showed that in some cases human language understanding exhibits certain non-classical phenomena [18, 140] and ambiguities [19], enabling quantum probabilities to serve as a suitable framework for modelling human language. Recently, the quantum measurement postulate, in conjunction with a set of procedural steps to transform a classical model to its quantum analogue, has been simulated into the neural network modelling paradigm for NLP tasks [78, 139]. The quantum-probability networks have not only achieved a SOTA performance but also demonstrated a high-level explainability in terms of model

transparency due to their theoretical root on the well-established quantum physics meanings. A more detailed description of related work in this direction has been presented in chapter 3. Nevertheless, existing approaches treated the interactions among quantum states as either a classical mixture of states [78, 139] or a separable product of states (chapter 4), which cannot fully exploit the potentials of quantum probabilities in modelling the entanglement, i.e., non-separability of multiple modalities. The expressiveness of quantum probabilities goes beyond classical correlations, describing joint probability distributions that cannot be decomposed into the tensor product of the individual ones. Such non-separability has been shown in cognitive science as a fundamental feature of human decision making under uncertainty, and it complies with the more general quantum probability theory only.

In line with this observation, we propose a joint quantum probabilistic neural network which captures non-classical correlations among distinct modalities. In particular, we transform the pre-trained real-valued uni-modal embeddings into pure quantum states of complex values. The way in which modality states then interacts with each other makes them non-separable, i.e., entangled. The main difference of this work from the previous probabilistic neural network approaches resides in the issues of *contextuality*, i.e., a modality activates multiple emotion senses in the context of the previous utterances and other modalities, and *non-separability*, i.e., a modality cannot be separated from the rest of modalities occurred concurrently. We have evaluated the proposed model on two SOTA benchmarking datasets for video emotion recognition in conversations, namely, IEMOCAP [27] and MELD [115]. Numerical experiments show that the model significantly optimizes performance compared to a wide variety of SOTA approaches. Moreover, the degree of non-separability of entangled states improves the post-hoc interpretability as well.

To summarize, this work makes the following contributions:

- A transparent and joint quantum probabilistic neural model, which models cross-modal correlations as non-separable, which to our best knowledge, no existing models in the current literature have taken into account.
- In contrast to previous work, the model considers both context modelling, in the form of preceding utterances, and multimodal fusion in the form of modality interactions, into a unified framework.
- The proposed model architecture supports multiparty conversations without requiring artificial expansion.
- The model achieves an improved performance, as compared to various SOTA approaches, for video emotion recognition in conversations.
- The degree of non-separability of modalities unearths useful and explainable knowledge about the way distinct modalities interact with each other.

5.2.2 Task Formulation

In general, this study is concerned with the task of inferring emotion recognition labels (e.g., happy, angry, sad, etc.) of constituent utterances u_1, u_2, ..., u_n in a video conversation. Each utterance u_k is associated with linguistic, visual, and acoustic content $u_k = \{u_{k,l}, u_{k,v}, u_{k,a}\}$ and uttered by an associated party p_s, where s is the index of the corresponding party. Essentially, the objective is to establish a mapping function that maps each constituent utterance to its corresponding emotion recognition label. Therefore, this is a multi-class classification task which involves two challenging issues: conversational context and non-separability of multiple modalities that dynamically interact with each other to determine the final emotion label.

This section presents a contextual and non-separable modality fusion strategy to address both issues in a unified neural network framework driven by quantum entanglement.

5.2.3 Model

A quantum system requires a set of procedural steps that convert information to its quantum analogue, evolve it through time evolution, and then perform measurements upon the system to make predictions. In line with this, we propose a transparent end-to-end quantum probabilistic neural model, namely, contextual Entanglement-driven Fusion Neural Network (c-EFNN).

General Architecture

The architecture of c-EFNN is illustrated in Figure 5.6. In particular, we have deployed an intermediate fusion strategy [119], which is a flexible multimodal fusion approach, into a neural network modelling paradigm that incorporates QT-inspired complex representation of information, composite quantum system and entanglement to capture the non-separability of modalities, quantum measurement for abstract feature extraction, and quantum evolution to model the contextuality over consecutive utterances and modalities.

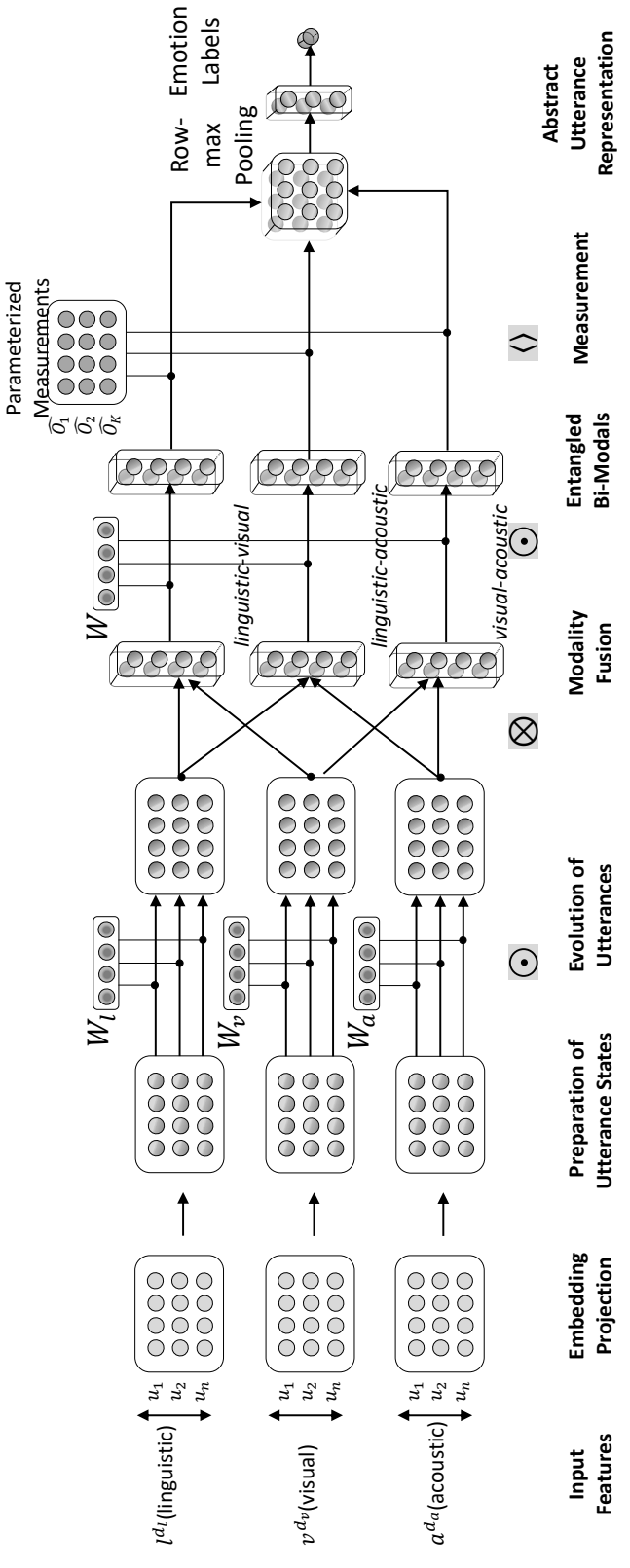


Figure 5.6: Contextual Entanglement-driven Fusion Neural Network (c-EFNN) architecture. The symbol $\odot$ stands for element-wise vector product, $\otimes$ the tensor product of vectors, and $\langle\rangle$ the inner product of vectors. Different shades imply transformations. The dimension of vector might vary over the procedural steps.

Specifically, c-EFNN takes multimodal information, i.e., linguistic, visual, and acoustic, for a sequence of utterances and feeds it into three separated neural branches, one for each modality (see Figure 5.6). After the preparation of utterance states, weight vectors are exploited to capture contextual information in the form of preceding utterances (see Figure 5.6, Evolution of Utterances step). Then, for each utterance, we operate a pairwise fusion of modalities, i.e., *linguistic-visual*, *linguistic-acoustic*, and *visual-acoustic*, via the tensor product of bi-modals (any of two modalities). Another weight vector models correlations within the bi-modal tensor-based representations (see Figure 5.6, Entangled Bi-Modals step). A set of parameterized measurements map the complex-valued representation to a real-valued high-level representation via the quantum measurement postulate. Finally, a row-max pooling operator followed by a fully connected layer passed to a *softmax* function for classification. In the remaining part of the section, we have elaborated on the methodology of the procedural steps.

Preparation of States

In this work, the utterances are modelled as uni-modal pure quantum states into modality-specific Hilbert spaces $\mathbb{H}_m$, where $m \in \{l, v, a\}$, for linguistic, visual, and audio modalities. In line with previous work [78, 139], we consider the exponential form of complex numbers to express quantum states: $z = re^{i\theta}$, where amplitude r is a real non-negative coefficient, phase $\theta \in [0, 2\pi)$, and i is imaginary number satisfying $i^2 = -1$.

Then, according to Equation 3.1, the modality-specific pure state of the k^{th} utterance $|u_{k,m}\rangle$, in short $|u_m\rangle$ to avoid clattering subscripts, could generally be expressed as

$$\begin{aligned} |u_m\rangle &= [r_{1,m}e^{i\theta_{1,m}}, r_{2,m}e^{i\theta_{2,m}}, ..., r_{d,m}e^{i\theta_{d,m}}]^T \\ &= [r_{1,m}, r_{2,m}, ..., r_{d,m}]^T \odot e^{i[\theta_{1,m},\theta_{2,m},...,\theta_{d,m}]^T} \end{aligned} \tag{5.13}$$

where d is the dimension of modality features and $\odot$ refers to element-wise vector product.

In Equation 5.13, the first vector, i.e., $r_m = [r_{1,m}, r_{2,m}, ..., r_{d,m}]^T$, corresponds to amplitudes, where the moduli r is a real-valued vector of unit length. To construct amplitudes, we transform pre-trained real-valued embeddings to their quantum analogues as follows. Suppose the input utterance-level features are $l \in \mathbb{R}^{d_l}$, $v \in \mathbb{R}^{d_v}$, $a \in \mathbb{R}^{d_a}$, for linguistic, visual, and acoustic modalities respectively. At the outset, we project the input features into the same dimension d via a fully connected layer with Rectified Linear Unit (ReLU) as the activation function, to ensure all elements $\{r_{i,m}\}_{i=1}^{d}$ are non-negative : $\hat{m} = ReLU(W_m m + b_m)$, where $m \in \{l, v, a\}$. Then, we normalized the outputs to create vectors of unit length: $r_m = \frac{\hat{m}}{||\hat{m}||^2}$.

The second vector in Equation 5.13, i.e., $\theta = [\theta_{1,m}, \theta_{2,m}, ..., \theta_{d,m}]^T$, is also real-valued, with all its elements in $[0, 2\pi]$. The assignment of the phases θ is an open research question. In this work, we enable the utterances to carry *temporal information*, i.e., the position of utterances in a conversation, and *speaker information*, i.e., the index of corresponding party, in the phase part. The phase θ of

the k^{th} utterance is hence calculated by

$$\theta = \theta(k, s) = f_{pe}(k) + f_{se}(s), \tag{5.14}$$

where $f_{pe}(k)$ defines a map $f_{pe} : \mathbb{N} \to \mathbb{R}^d$ from a discrete position index to a $d-$dimensional real-valued vector, and $f_{se}(s)$ is a map $f_{se} : \mathbb{N} \to \mathbb{R}^d$ from a discrete index s of the corresponding party to a $d-$dimensional real-valued vector. To constrain $\theta \in [0, 2\pi]$ during training, we transform the real values to a uniform distribution $U[0, 2\pi]$. In this way, the non-recurrent architecture of c-EFNN not only captures the sequential information of utterances, but also handles the speaker emotion dependencies, which has been a major issue for conversational emotion recognition [116].

Time Evolution

In this work, we aim to learn the join-association between multiple utterances and modalities via the concept of evolution, and to emphasize on the contributing modalities via the concept of non-separability. Thus, we model both the time evolution of consecutive utterances and distinct modalities to track the dynamics of emotional states in a conversation. However, the main difference of the evolution of consecutive utterances in relation to the evolution of distinct modalities lies in the way the quantum states of the later interact so that they are not to be separable.

Time Evolution of Consecutive Utterances After the transformation of feature inputs to quantum states into uni-modal Hilbert spaces, modality-specific complex-valued neural layers are injected (one each for linguistic, visual, and acoustic), to update the quantum states taking into account preceding utterances (see Figure 5.6, Evolution step). For the k^{th} utterance state $|u_{k,m}\rangle$, where $m \in \{l, v, a\}$, the transformation induced by the neural layer is given by

$$|\widehat{u_{k,m}}\rangle = W_m \odot |u_{k,m}\rangle\,, \tag{5.15}$$

which can be considered as weighted linear transformation layer, where $W_m \in \mathbb{R}^d$ is the modality-specific weight vector, and $\odot$ stands for element-wise vector product. The output is an unnormalized vector $|\widehat{u_{k,m}}\rangle$, which is then normalized to get a vector of unit length, i.e., a valid quantum state: $|\widehat{u_{k,m}}\rangle = \frac{\widehat{u_{k,m}}}{||\widehat{u_{k,m}}||^2}$, in short $|\widehat{u_{k,m}}\rangle = |u_m\rangle$. The W_m is a joint operator, which interacts with the consecutive utterance states. From the quantum point of view, this means that Equation 5.15 is a valid approximation of Equation 3.15 [14], i.e., W_m can be realized as a unitary operator U. Despite W_m acts as a quantum Hamiltonian control which interacts with individual utterance quantum states, the entanglement may not be present between utterances, since they share the same specific-modality Hilbert space. From the representation learning point of view, W_m captures the contextual information in a conversation [149].

Entanglement and Time Evolution of Distinct Modalities The updated uni-modal utterance states are fed into the modality fusion component (see Figure 5.6). In particular, we deploy a fusion module which takes the utterance states of *pairwise* modalities, i.e., *linguistic-visual*, *linguistic-acoustic*, *visual-acoustic*. For each pairwise of states, a composite separable state is created by computing the tensor product of them. The composite separable state is defined on a d^2-dimensional joint space $\mathbb{H}_{m_1,m_2} := \otimes_2(H_m)_2$ and formulated as

$$|u_{m_1,m_2}\rangle = |u_{m_1}\rangle \otimes |u_{m_2}\rangle\,, \tag{5.16}$$

where $m1, m2$ any of two modalities, and $\otimes$ defines the outer product of two states.

Then, another complex-valued neural layer W is connected to induce interactions of pairwise modalities (see Figure 5.6), as follows:

$$|\widehat{u_{m_1,m_2}}\rangle = W \odot |u_{m_1,m_2}\rangle\,, \tag{5.17}$$

the output of which is an unnormalized vector. Likewise, we normalize the output to get a unit vector in $\mathbb{H}_{m_1,m_2}$, which corresponded to a pure quantum state: $|\widehat{u_{m1,m2}}\rangle = \frac{\widehat{u_{m_1,m_2}}}{||\widehat{u_{m_1,m_2}}||^2}$, in short $|\widehat{u_{m_1,m_2}}\rangle = |u_{m_1,m_2}\rangle$. Throughout the update process of utterance states, W_m has acted as a quantum Hamiltonian control, i.e., a joint operator on multiple states, on the same Hilbert space $\mathbb{H}_m$. Conversely, in this case, W acts as a quantum Hamiltonian control on different Hilbert spaces, i.e., $\mathbb{H}_{m_1}$, $\mathbb{H}_{m_2}$, and ***entanglement*** is hence generated after the transformation. This means that the output after the transformation cannot be written in the decomposable form of Equation 3.12, thus giving the potential to capture non-classical correlations across pairwise modalities.

Measurement

The measurement component acts upon n sets of three non-separable pairwise modalities, one set for each utterance, to identify the discriminating information to emotion classification. In particular, a set of parameterized measurements $\{O_k\}_{k=1}^{K}$ are performed to the set of non-separable pairwise modalities (see Figure 5.6), generating a sequence of positive scalars via

$$P(k) = |\langle O_k|u_{m_1,m_2}\rangle|^2, \tag{5.18}$$

where m_1, m_2 are any pair of modalities and each O_k represents an abstract emotion concept. The output is a $K \times 3$ matrix of positive real values produced by measurement. Each value corresponds to the likelihood of a non-separable pairwise modality state collapsing to a basis state O_k, which is effectively a basis context representing abstract emotion concepts. Note that the measurement component can be thought of as a dictionary learning approach [83]. Then a row-wise maximum pooling operator is conducted to cascade the three sequences of abstract emotions into one high-level

utterance representation (see Figure 5.6). Finally, the high-level representation is passed to a fully connected layer followed by a softmax classifier.

5.2.4 Experiments

Datasets

We performed experiments on two widely used benchmarking datasets in conversational emotion detection: IECAMOCAP [27] and MELD [115]. Table 5.1 summarizes the statistics of both datasets in terms of trainning, validation and test sets. IEMOCAP consists of videos of dyadic conversations between pairs of 10 speakers. The videos are segmented into utterances with annotations of fine-grained emotion categories. In this work, we consider six such categories for the classification task: *anger*, *happiness*, *sadness*, *neutral*, *excitement*, and *frustration*. The training and validations sets are curated using the first 8 speakers. MELD is a multiparty dialog emotion classification dataset. It has more than 1400 dialogues and 13000 utterances from the Friends TV series. Each utterance in every dialog is annotated with emotions. We considered the following emotion categories: *anger*, *sadness*, *joy*, *surprise*, and *neutral*. Figure 5.7 shows an example of multimodal conversation for MELD task.

	Dialogue			Utterances		
Dataset	**Train**	**Val**	**Test**	**Train**	**Val**	**Test**
IECAMOCAP	96	24	31	6808	1702	1623
MELD	1039	114	280	9989	1109	2610

Table 5.1: Training, validation and test data distribution in the datasets.

For a fair comparison, we used the publicly available[3] pre-trained utterance features provided by the authors of DialogueRNN [92], a SOTA model. Each utterance in IEMOCAP had a 100-dimensional textual feature vector, a 512-dimensional visual feature vector and 100-dimensional acoustic feature vector. On the other hand, for each utterance in MELD, 600-dim textual features and 300-dim acoustic features were used[4]. For visual features, the 3D-CNN [65] is adopted to extract abstract representations from raw video data. The 3D-CNN network conducts convolution operation on the height and width for a video frame and across the time domain to be capable of learning the temporal relationships between consecutive frames. The dimension of features is 300.

[3]https://github.com/declare-lab/conv-emotion
[4]The pre-trained visual feature is not publicly available for MELD

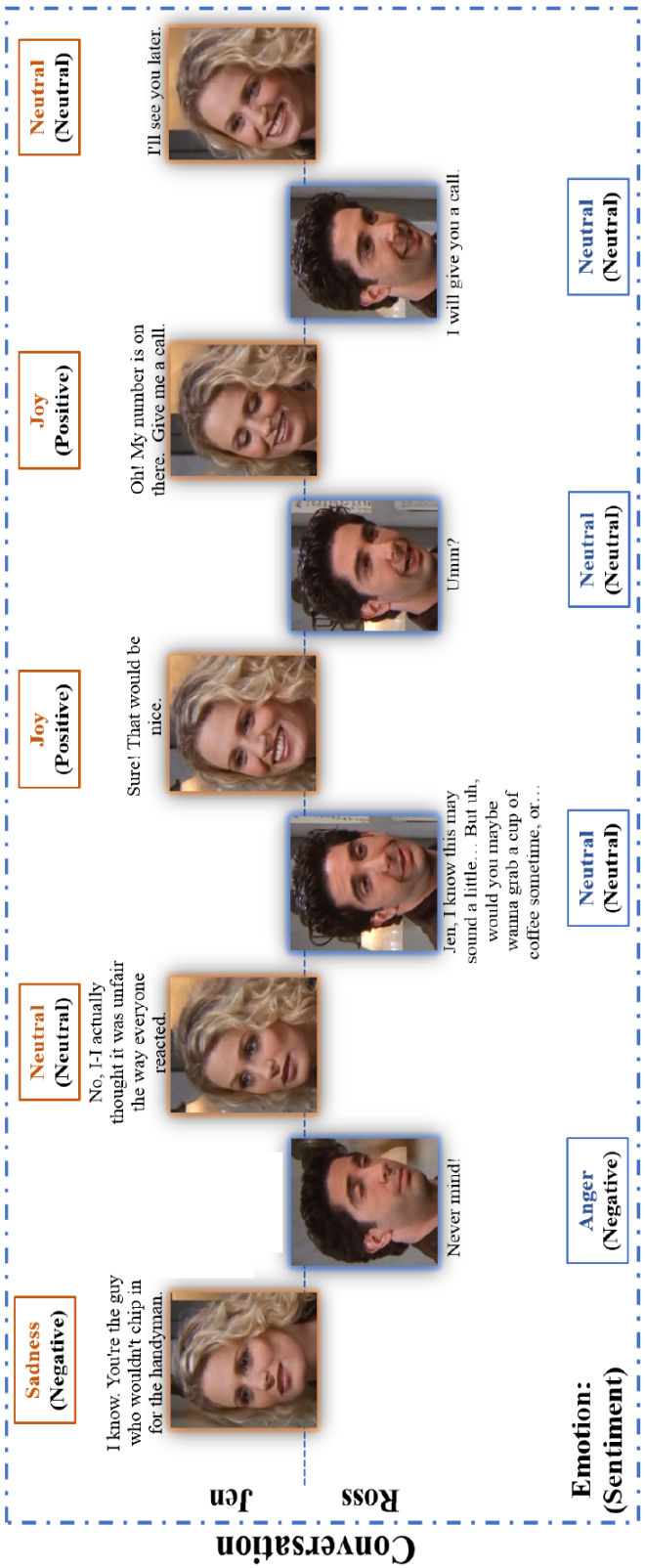


Figure 5.7: An example of the multimodal conversation. The task is to predict the emotion of each utterance in the conversation, considering preceding utterances as well.

Baselines

We compared the proposed c-EFNN model with a great variety of the SOTA monologue, dialogue and quantum-inspired neural approaches in affective computing.

Monologue Models To adopt the monologue models to the conversational setting, we changed the word-level input features to utterance-level input features, and an output was yielded at each timestamp. We replicated the following monologue baselines: a) **TFN** [151] is a tensor-based neural network which captures uni-modal, bi-modal, and tri-modal interactions across distinct modalities via a multi-dimensional tensor, b) **MFN** [152] exploits a hybrid memory cell, constructed from the hidden units of each modality at the previous timestamp, which acts as an additional input of the next timestamp, and c) **MulT** [126] consists of pairwise cross-modal transformers, the outputs of which are concatenated to build a multimodal embedding for each utterance.

Dialogue Models We also replicated the following dialogue models: a) **c-LSTM** [113] uses a Bi-directional Long Short-Term Memory (LSTM) for each modality, to capture the contextual content from surrounding utterances. The context-aware utterance representations are then used as inputs into another Bi-directional LSTM for emotion classification, b) **c-GRU+Att**[47] is a variant of c-LSTM, by replacing LSTM with Gated Recurrent Unit (GRU) and applying an attention mechanism component to capture cross-modal interactions, c) **CMN** [60] uses two distinct GRUs for two parties, i.e., a speaker and a listener. The utterance representation is obtained by feeding the current utterance as a query to two distinct memory networks for both parties, d) **ICON** [59] is an extension of CMN, connecting outputs of individual speaker GRUs in CMN using another GRU, which acts as a memory to track the overall conversational flow and e) **DialogueRNN** [92] is a recurrent network that uses two GRUs to track individual speaker states and global context during the conversation. Further, another GRU is employed to track emotional state through the conversation.

Quantum-inspired Models We finally compare with **QMF** [73] introduced in chapter 4.

For evaluation, we chose average accuracy and F_1 scores over all emotion categories. Besides, for a fine-grained understanding of performance, precision, recall, and $F1$ scores were calculated for each emotion category.

Experimental Settings

A grid search for the best hyper-parameters was conducted for all models. At each search, models were trained for 100 epochs, and the model with the lowest validation loss was chosen. Out of 50 searches, the model with the highest average F1 score on the test set was taken to compute the performance values. CMN [60], ICON [59], and DialogueRNN [92] were excluded from the MELD experiment, since they cannot support multiparty conversations [48].

The parameters in the proposed c-EFNN model were determined by the set of hyper-parameters $\Theta = \{N, D, K, S\}$, where N is the number of utterances in a conversation, D is the embedding

dimension of input features after projection layers, K is the number of measurement vectors, and S is the number of speakers in a conversation. For both datasets, we searched over a parameter pool : $D \in \{100, 120, 140, 160, 180, 200\}$, size of last hidden layer $\in \{32, 48, 64, 80\}$, dropout rate for the last hidden layer $\in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$, learning rate $\in \{0.001, 0.002, 0.005, 0.008\}$. The batch size varied in proportion to the dataset scale, i.e., batch size$\in \{24, 48, 96\}$ for MELD and batch size $\in \{4, 8, 16\}$ for IEMOCAP.

We trained c-EFNN by feeding the real and imaginary parts of the complex-valued layers as different input parts and simulated complex operations using real values [125]. Indeed, any complex function $f : \mathbb{C}^n \to \mathbb{C}^n$ can be represented as $f(a + ib) = a(a, b) + ib(a, b)$, where $a, b : \mathbb{R}^n \to \mathbb{R}^n$ [8]. c-EFNN was hence trained via the backpropagation algorithm. Measurements were initialized from standard normal distributions. All the parameters were trainable with respect to cross-entropy loss defined on the extracted features. We chose Adam as the optimization algorithm.

Performance Analysis

Table 5.2 shows the comparison results between c-EFNN and various monologue, dialogue, and quantum-inspired approaches for the IEMOCAP task. For most of the strategies, the emotion category about happiness is the most challenging due to under-fitting. Indeed, this emotion category has fewer samples compared to other emotion categories in IEMOCAP. We notice that almost all baseline strategies struggled with neutral utterances, although neutral emotion category has the most samples in IEMOCAP. Indeed, in contrast to QT-based modelling, which tackles ambiguities in content, the traditional approaches in affective computing cannot effectively cope with utterances when their context is uninformative or ambiguous [49]. In general, c-EFNN achieves the best precision or recall for most of the emotion categories in IEMOCAP. Overall, c-EFNN attains an increased average accuracy of 64.46% as compared to 62.30% of DialogueRNN (see Table 5.2). That is a significant improvement of 3.5% (t-test$<$.05). c-EFNN also yields an increased average F_1 score of 64.30% as compared to 59.60% of c-GRU+Att, i.e., a 7.3% improvement (t-test$<$.05).

	Happy			Sad			Neutral			Angry			Excited			Frustrated			Average	
Model	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Accuracy	F1
Monologue																				
TFN [151]	0.4150	0.3155	0.3585	0.7250	0.6170	0.6667	0.5080	0.5505	0.5284	0.6280	0.6220	0.6250	0.6310	0.5570	0.5917	0.5540	0.6660	0.6049	0.5710	0.5725
MFN [152]	0.4210	0.3190	0.363	0.7420	0.6230	0.6773	0.5220	0.5580	0.5394	0.6250	0.6180	0.6215	0.6530	0.5720	0.6098	0.5600	0.6692	0.6097	0.5810	0.5801
MulT [126]	0.4370	**0.3850**	0.4094	0.7320	0.6910	0.7109	0.5180	0.5210	0.5195	0.6603	0.5720	0.6103	0.6900	0.6032	0.6437	**0.5590**	0.6810	0.6140	0.5964	0.5946
Dialogue																				
c-LSTM [113]	0.3870	0.3290	0.3557	0.7810	0.5850	0.6689	0.5590	0.4830	0.5182	0.5510	0.7390	0.6313	0.6730	0.6610	0.6669	0.5530	0.6690	0.6050	0.5890	0.5840
c-GRU+Att [47]	0.4310	0.3710	0.3980	0.7305	0.6910	0.7102	0.5280	0.5290	0.5285	0.6630	0.5770	0.6170	**0.6910**	0.6145	0.6505	0.5555	0.6810	0.6116	0.5970	0.5960
CMN [60]	0.4270	0.3810	0.4027	0.7770	0.5405	0.6375	0.5500	0.5830	0.5660	0.6105	**0.6360**	0.6230	0.6740	0.5990	0.6343	0.5680	0.7115	0.6317	0.5956	0.5925
ICON [59]	0.3840	0.3310	0.3555	0.7570	0.5340	0.6262	0.5170	**0.6000**	0.5554	0.6350	0.5990	0.6165	0.6330	0.5410	0.5834	0.5550	0.6705	0.6043	0.5780	0.5668
DialogueRNN [92]	0.8610	0.1520	0.2584	0.8170	0.6995	0.7537	0.6205	0.4850	0.5444	0.6450	0.5300	0.5819	0.6510	**0.8210**	**0.7262**	0.5290	0.7910	0.6340	0.6230	0.5070
Quantum																				
QMF [73]	0.4070	0.3710	0.3882	0.7290	0.6390	0.6810	0.5470	0.5550	0.5510	0.6490	0.5970	0.6220	0.6500	0.6710	0.6603	0.5505	0.6982	0.6156	0.5970	0.5880
c-EFNN	**0.8730**	0.3350	**0.4842**	**0.8280**	**0.7095**	**0.7642**	**0.6290**	0.5520	**0.5880**	**0.6705**	0.6210	**0.6448**	0.6895	0.6570	0.6729	0.5410	**0.7960**	**0.6442**	**0.6446** (↑ 3.5%)	**0.6430** (↑ 7.3%)

Table 5.2: Effectiveness of c-EFNN on IEMOCAP. Best results are highlighted in bold. Numbers in parentheses indicate relative percentage improvement over the next best model.

	Sad			Neutral			Angry			Surprise			Joy			Average	
Model	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Accuracy	F1
Monologue																	
TFN [151]	0.2500	0.1355	0.1757	0.7210	0.7720	0.7456	0.3470	0.3995	0.3714	0.4120	0.5710	0.4786	0.5170	0.4770	0.4961	0.5760	0.5434
MFN [152]	0.2890	0.2135	**0.2456**	0.7152	**0.8410**	0.7730	**0.4690**	0.3920	0.4270	**0.5890**	0.4055	0.4803	0.4830	**0.5980**	**0.5344**	0.6005	0.5820
MulT [126]	0.3680	0.1690	0.2316	0.7210	0.8195	0.7671	0.4310	0.5130	0.4684	0.4990	0.4770	0.4877	0.5180	0.5325	0.5251	0.6075	0.5859
Dialogue																	
c-LSTM [113]	0.3317	**0.1912**	0.2426	0.7250	0.7980	0.7598	0.4330	0.4510	0.4418	0.4590	0.5582	**0.5038**	0.5230	0.5372	0.5300	0.6065	0.5856
c-GRU+Att [47]	**0.4250**	0.1320	0.2014	0.7285	0.8020	0.7635	0.4470	0.3915	0.4174	0.4235	**0.6210**	0.5036	0.4955	0.5475	0.5202	0.5995	0.5810
Quantum																	
QMF [73]	0.2510	0.1270	0.1686	0.7120	0.8290	0.7660	0.4250	0.4450	0.4348	0.4480	0.5320	0.4864	**0.5350**	0.5170	0.5258	0.6045	0.5816
c-EFNN	0.3740	0.1740	0.2375	**0.7325**	0.8335	**0.7797**	0.4395	**0.5205**	**0.4766**	0.5070	0.4850	0.4957	0.5290	0.5400	**0.5344**	**0.6310** (↑ 3.7%)	**0.5944** (↑ 1.4%)

Table 5.3: Effectiveness of c-EFNN on MELD. Best results are highlighted in bold. Numbers in parentheses indicate relative percentage improvement over the next best model.

Table 5.3 presents the results for the MELD task. Unlike IEMOCAP task, the baseline approaches generally achieve reasonably high performance for neutral utterances due to the abundance of neural utterances in MELD. Indeed, half of the samples in MELD belong to neutral emotion category, i.e., 4750 neutral utterances out of 9450 in the training set. In contrast, the baseline approaches generally show a relatively low performance for the other categories, e.g., in most cases, the precision was lower than 50%. We suspect that this is due to the skewness of data, in terms of the numbers of samples between the neutral emotion category and the rest of emotion categories in MELD. Among the baselines, MulT, which is a monologue-based approach, achieves the best performance. However, c-EFNN gains an increased average accuracy of 63.10% as compared to 60.75% of MulT, which is a significant improvement of 3.7% (t-test$<$.05). Finally, c-EFNN again attains an improved F_1 score of 59.44% as compared to 58.59% of MulT.

In summary, c-EFNN significantly outperforms the baselines for both IEMOCAP and MELD tasks. The analysis of results has shown that c-EFNN is capable of coping with ambiguous utterances and skewed datasets.

Ablation Test

We also carried out an ablation test on MELD to investigate the effect of introduced quantum components. In particular, to examine the effectiveness of the time evolution-based modelling of consecutive utterances, we replaced the component with GRU layers (a.k.a. *c-EFNN-gru*). The outputs were normalized to get vectors of unit length. We investigated the impact of non-separable modalities, introducing two other variants of c-EFNN, after removing the weight vector W (see Figure 5.6): a) *c-EFNN-tensor* fuses all modalities into a unified tensor-based representation; and b) *c-EFNN-con* concatenates all modalities into a vector representation. Then, the outputs interacted with the measurement component. We also considered the impact of speaker influences by initializing phases from standard normal distributions (a.k.a. *c-EFNN-rand*). We finally replaced measurements with a convolutional neural network (CNN), whereby the K filters of CNN served as K measurements to investigate the impact of the measurement component (a.k.a. *c-EFNN-cnn*).

The results of the ablation test, illustrated in Table 5.4, show that each component plays a crucial role in the c-EFNN. In particular, the comparison with c-EFNN-rand shows the effectiveness of modelling speaker influences into the phase part of complex-valued representations, while the increase in performance over c-EFNN-tensor and c-EFNN-con reveals the superiority of encoding the non-classical correlations (i.e., entanglements) between modalities. Moreover, the comparison with c-EFNN-cnn shows the usefulness of trainable measurements. Finally, c-EFNN-gru turns out to be a less appropriate way to construct updated utterance states.

	Average	
Variant	Accuracy	$F1$
c-EFNN-gru	0.6235 (↓ 1.2%)	0.5860 (↓ 1.4%)
c-EFNN-tensor	0.6190 (↓ 1.9%)	0.5810 (↓ 2.2%)
c-EFNN-con	0.6140 (↓ 2.7%)	0.5760 (↓ 3.0%)
c-EFNN-rand	0.6220 (↓ 1.4%)	0.5830 (↓ 1.9%)
c-EFNN-cnn	0.6170 (↓ 2.2%)	0.5805 (↓ 2.3%)

Table 5.4: Ablation test of c-EFNN on MELD. Values in parentheses are the relative percentage differences from c-EFNN.

Post-hoc Interpretability

Further, we evaluated the post-hoc interpretability by investigating bi-modal correlations within composite utterance states after the modality context interaction. In particular, according to Equation 3.11, we calculated the degree of quantum entanglement for bipartite composite utterance states of linguistic and visual modalities.

Linguistic	Visual	Emotion
It turns out, we can't do it. Monica has to work.		Joy
Oh yeah, that's right!		Anger
You can't fire him and dump him the same day, he'll kill himself.		Sadness

Table 5.5: Selected most entangled linguistic-visual modalities on MELD.

Table 5.5 and Table 5.6 illustrate examples of the most and least entangled linguistic-visual modalities according to entanglement entropy. The most entangled pairs are those that one of two modalities is ambiguous or uninformative. For example, in Table 5.5, the linguistic content of the two first utterances is ambiguous, while the visual content of the third utterance is uninformative. By contrast, when the context of both modalities is informative, unambiguous, and simultaneously present (see Table 5.6), the entanglement entropy is close to zero. In those cases, the composite representation is separable, and there is no need to exploit the quantum probabilistic interpretation.

However, through the concept of non-separability, c-EFNN is able to capture both separable and non-separable bi-modal interactions, as a generalization of existing probabilistic modality fusion approaches. This attribute is the core reason that c-EFNN has achieved improved performance.

Linguistic	Visual	Emotion
I really think I might kill someone tonight!		Anger
I want to start drinking in the morning. Don't say that I don't have goals!		Anger

Table 5.6: Selected examples of least entangled linguistic-visual modalities on MELD.

We also investigated the number of effective degrees of freedom contributing to the entanglement of bi-modals. In particular, we calculated the average score of Schmidt numbers according to Equation 3.13 for pairwise modalities. Table 5.7 shows the average scores for linguistic-visual, linguistic-acoustic, and visual-acoustic modalities on IEMOCAP and MELD, respectively. When values are greater to 1, the bipartite modality states are non-separable, i.e., they are entangled. In particular, the bigger the average score of Schmidt numbers is, the higher the degree of entanglement.

Bi-modals	IEMOCAP	MELD
Linguistic-Visual	52.6	25.2
Linguistic-Acoustic	49.8	23.2
Visual-Acoustic	1.8	$\sim$1

Table 5.7: Average Schmidt scores of bi-modals on IEMOCAP and MELD tasks

The results show that entanglement is substantially present for linguistic-visual and linguistic-acoustic bi-modals. This implies that the linguistic modality plays a predominant role in determining the emotional state of utterances. That is, linguistic modality is more informative compared to ambiguous visual and acoustic modalities. In particular, linguistic modality acts as a context for visual and acoustic modalities, and as soon as modalities are simultaneously present, the utterance's emotional state becomes apparent. We notice that the degree of entanglement is higher for linguistic-visual bi-modals as compared to linguistic-acoustic. This implies that visual modality is less informative than acoustic, and there are hence more changes for visual modality to be entangled with linguistic, due to its ambiguous content.

In contrast, there are only a few non-separable visual-acoustic modality states on IEMOCAP, while the visual-acoustic modality states are separable on MELD task. We attribute this result

to the uninformative content of visual and acoustic modalities. Even though both modalities are simultaneously present, the ambiguous content of visual modality can not act as context for the ambiguous acoustic modality as well, to clarify the overall emotional state, and vice versa.

We finally observed that the effective degree of freedom contributing to the entanglement of bi-modals is higher on IEMOCAP as compared to MELD. We suspect this is because IEMOCAP is a small dataset, and the degree of ambiguity is high. We also speculate that the difference is due to the dimensions of sub-spaces. Roughly speaking, the smaller the dimension of sub-spaces are, the bigger the Schmidt number [31]. For our experiments, the embedding dimensions of input features were 100 and 160 for IEMOCAP and MELD, respectively.

5.2.5 Section Conclusions

In this work, we have introduced a transparent and joint quantum probabilistic neural model for video conversational emotion recognition, which addresses context modelling and multimodal fusion challenges into a unified framework. The model is based on the concept of non-separability to fuse bi-modals, capturing classical and non-classical correlations between modalities. Our experiments on both benchmarking datasets demonstrate the effectiveness of encoding bi-modal information in the form of non-classical correlations. Besides, non-classical correlations were quantified by appropriate measures, which optimized post-hoc interpretability.

5.3 An Entanglement-driven Fusion Neural Network for Video Sentiment Analysis

In this section, we present a variance of c-EFNN model introduced in Section 5.2. In particular, this model disregards context modelling in the form of preceding utterances and emphasizes on the contributing modalities via the concept of non-separability. Extensive empirical evaluation is carried out on two large-scale benchmarking datasets for video sentiment analysis. The model, namely EFNN, achieves significant improvements over a wide range of SOTA baselines and increased post-hoc interpretability. The task is an utterance-level sentiment analysis as discussed in Chapter 2, Section 2.2.1.

5.3.1 Model

The architecture of EFNN is illustrated in Figure 5.8. The main difference between c-EFNN (Figure 5.6) and EFNN (Figure 5.8) is that EFNN does not include the Evolution of Utterances component, which captures contextual information in the form of preceding utterances. Another crucial change is that c-EFNN receives utterance-level inputs from a dialogue, whilst EFNN gets word-level features from sentences.

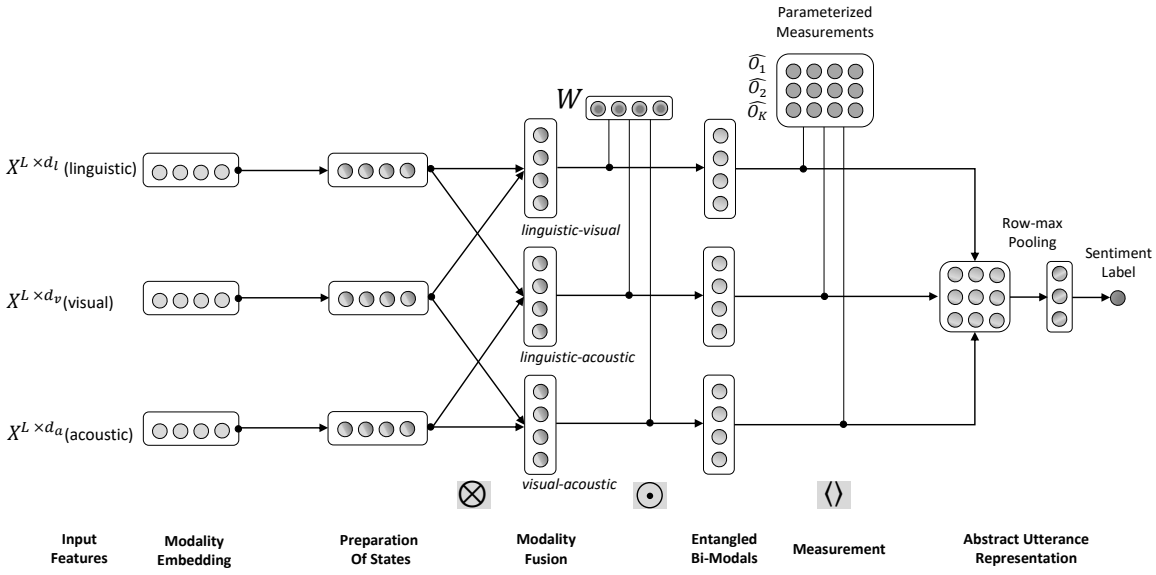


Figure 5.8: Entanglement-driven Fusion Neural Network (EFNN) architecture. The symbol $\otimes$ stands for the tensor product of vectors, $\odot$ the element-wise vector product, and $\langle\rangle$ the inner product of vectors. Different shades imply transformations. The dimension of vector might vary over the procedural steps.

Preparation of States

In this work, each utterance is modelled as a uni-modal pure quantum state into modality-specific Hilbert spaces $\mathbb{H}_m$, where $m \in \{l, v, a\}$, for linguistic, visual and audio modalities. In line with previous works [78, 73, 139], we consider the exponential form of complex numbers to express quantum states: $z = re^{i\theta}$, where amplitude r is a real non-negative coefficient, phase $\theta \in [0, 2\pi)$, and i is imaginary number satisfying $i^2 = -1$.

Then, according to Equation 3.1, the modality-specific pure state of an utterance $|u_m\rangle$ could generally be expressed by the following modulus-augment form:

$$\begin{aligned} |u_m\rangle &= [r_{1,m}e^{i\theta_{1,m}}, r_{2,m}e^{i\theta_{2,m}}, ..., r_{d,m}e^{i\theta_{d,m}}]^T \\ &= [r_{1,m}, r_{2,m}, ..., r_{d,m}]^T \odot e^{i[\theta_{1,m},\theta_{2,m},...,\theta_{d,m}]^T} \end{aligned} \tag{5.19}$$

where d is the dimension of modality features and $\odot$ refers to element-wise vector product. In the modulus-argument form, any operation on the complex numbers will lead to a non-linear combination of the constituent moduli and arguments. This implies that a non-linear feature combination is inherently produced when we assign Equation 5.19 with linguistic, visual, and acoustic features.

In Equation 5.19, the first vector, i.e., $r_m = [r_{1,m}, r_{2,m}, ..., r_{d,m}]^T$, corresponds to amplitudes, where the moduli r is a real-valued vector of unit length. To construct amplitudes, we transform

the input real-valued features to their quantum analogues as follows. Suppose the input word-level features are $X^l \in \mathbb{R}^{L\times d_l}$, $X^v \in \mathbb{R}^{L\times d_v}$, $X^a \in \mathbb{R}^{L\times d_a}$, where d_l, d_v, d_a represents feature dimensions for linguistic, visual, and acoustic modalities respectively, and L is the sequence length, i.e., total number of words in an utterance. At the outset, we project the input features into the same dimension d via convolutional neural networks [70] from the respective input features with Rectified Linear Unit (ReLU) as the activation function in the last hidden layer, to ensure all elements $\{r_{i,m}\}_{i=1}^{d}$ are non-negative: $\hat{m} = ReLU(CNN_m(X^m)) \in \mathbb{R}^d$, where $m \in \{l, v, a\}$. Despite the projection of modalities into a common space, the convolutional neural networks CNN_m capture local structure of words in an utterance. Then, we normalize the outputs to create vectors of unit length: $r_m = \frac{\hat{m}}{||\hat{m}||^2}$.

The second vector in Equation 5.19, i.e., $\theta = [\theta_{1,m}, \theta_{2,m}, ..., \theta_{d,m}]^T$, is also real-valued, with all its elements in $[0, 2\pi]$. The assignment of the phases θ is an open research question. In this work, to enable each utterance to carry *temporal information*, i.e., we assign the position of words in a sentence to the phase part. With this way, we capture the global structure of words in an utterance. The phase θ is hence calculated by

$$\theta = \theta(k) = f_{pe}(k), \tag{5.20}$$

where $f_{pe}(k)$ defines a map $f_{pe} : \mathbb{N} \to \mathbb{R}^d$ from a discrete position index to a $d-$dimensional real-valued vector.

Entanglement-driven Modality Fusion

After the transformation of feature inputs to quantum states into uni-modal Hilbert spaces, we feed them into the modality fusion component (see Figure 5.8). In particular, we deploy a fusion module, which takes the utterance states of *pairwise* modalities, i.e., *linguistic-visual*, *linguistic-acoustic*, *visual-acoustic*. For each pairwise of states, a composite but separable state is created by computing the tensor product of them. The composite separable state is defined on a d^2-dimensional joint space $\mathbb{H}_{m_1,m_2} := \otimes_2(H_m)_2$ and formulated as

$$|u_{m_1,m_2}\rangle = |u_{m_1}\rangle \otimes |u_{m_2}\rangle\,, \tag{5.21}$$

where m_1, m_2 any of two modalities, and $\otimes$ defines the outer product of two states.

Then, a complex-valued neural layer W is injected to induce interactions of pairwise modalities (see Figure 5.8), as follows:

$$|\widehat{u_{m_1,m_2}}\rangle = W \odot |u_{m_1,m_2}\rangle\,, \tag{5.22}$$

where $W \in \mathbb{R}^{d^2}$ is a shared weight vector, and $\odot$ stands for element-wise vector product. The output is an unnormalized vector $|\widehat{u_{m_1,m_2}}\rangle$, which is then normalized to get a unit vector in $\mathbb{H}_{m_1,m_2} \in \mathbb{R}^{d^2}$, i.e., a valid quantum state: $|\widehat{u_{m_1,m_2}}\rangle = \frac{|\widehat{u_{m_1,m_2}}\rangle}{||\widehat{u_{m_1,m_2}}||^2}$, in short $|\widehat{u_{m_1,m_2}}\rangle = |u_{m_1,m_2}\rangle$.

5.3.2 Measurement

The measurement component acts upon the set of three non-separable pairwise modalities to identify the discriminating information for sentiment classification. In particular, a set of parameterized measurements $\{O_k\}_{k=1}^{K}$ are performed on the set of non-separable pairwise modalities (see Figure 5.8), generating a sequence of positive scalars for each pair of modalities via

$$P(k) = |\langle O_k | u_{m_1,m_2} \rangle|^2, \tag{5.23}$$

where m_1, m_2 are any pair of modalities and each O_k represents an abstract sentiment concept. The output is a $K \times 3$ matrix of positive real values produced by measurement. Each value corresponds to the likelihood of a non-separable pairwise modality state collapsing to a basis state O_k, which is in effect a basis context representing abstract sentiment concepts.

Then a row-wise maximum pooling operator is conducted to cascade the three sequences of abstract concepts into one high-level utterance representation (see Figure 5.8). Finally, the high-level representation is passed to a fully connected layer followed by a softmax classifier.

5.3.3 Experiments

Datesets and Evaluation Metrics

The experiments were conducted on two SOTA benchmarking video sentiment analysis datasets, namely CMU-MOSI [155] and CMU-MOSEI [157]. We have left the details about the datasets and feature extraction in chapter 2, Section 2.2.2.

To evaluate the effectiveness of our model on CMU-MOSI and CMU-MOSEI tasks, we adopted a series of evaluation performance metrics used in prior work [80, 126, 152, 157], including: binary accuracy (i.e., Acc_2 : positive sentiment if $values \geq 0$, and negative sentiment if $values < 0$), 7-class accuracy (i.e., Acc_7 : sentiment score classification in $Z \cap [-3, 3]$), $F1$ score, Mean Absolute Error (MAE) of the score, and the Pearson's correlation ($Corr$) between the model predictions and regression ground truth. For all the metrics, the higher values denote a better performance, except MAE where the lower values denote better performance.

Baselines

We compare our proposed **EFNN** model with the SOTA neural approaches for video sentiment analysis, replicating a variety of baseline Long Short-Term Memory (LSTM) [62], advanced LSTM, tensor, sequence-to-sequence, and quantum-inspired strategies.

Baseline LSTM: Early-Fusion LSTM (EF-LSTM) concatenates linguistic, visual, and acoustic features at each timestamp, and builds an LSTM to construct sentence-level multimodal representation. The last hidden state is taken and sequentially passed to two fully connected layers

to produce the output sentiment. **Late-Fusion LSTM (LF-LSTM)** builds LSTMs for linguistic, visual, and acoustic inputs separately, and concatenates the last hidden state of the three LSTMs as sentence-level multimodal representation. The concatenated hidden states are taken and sequentially passed to two fully connected layers to produce the output sentiment.

Advanced LSTM: Multi-Attention Recurrent Network (MARN) [153] captures cross-modal dynamics at each timestamp. A multi-attention block is built to construct a cross-modal representation, based on hidden states of the previous timestamp, and fed into the inputs of the current timestamp. The cross-modal representation and hidden states of the last timestamp are concatenated to form a multimodal sentence embedding, which is sequentially passed to two fully connected layers to produce the output sentiment. **Memory Fusion Network** (MFN) [152] is a memory fusion network that builds a multimodal gated memory component. The memory cell is updated along with the evolution of the hidden states of three unimodal LSTMs. The last memory cell is concatenated with the last hidden states of unimodal LSTMs to construct the multimodal sentence representation. Then, the multimodal representation is sequentially passed to two fully connected layers to produce the output sentiment. **Contextual GRU with Attention (c-GRU+Att)** [47] encodes linguistic, visual, and acoustic streams through three separate Bi-GRU layers followed by fully connected dense layers. Then, pairwise attentions are computed across all possible combinations of modalities. Finally, individual modalities and bi-modal attention pairs are concatenated to create the multimodal representation, used for final classification. c-GRU+Att makes predictions by applying a fully connected layer to each timestamp. In our experiments, since we did not consider preceding utterances, we extracted the last hidden state only and fit it to a fully connected layer to make predictions.

Tensor: Tensor Fusion Network (TFN) [151] explicitly models view-specific and cross-view dynamics by creating a multi-dimensional tensor that captures unimodal, bi-modal, and tri-modal interactions across linguistic, visual, and acoustic modalities. **Low-rank Multimodal Fusion (LMF)** [85] adopts the same approach as TFN to model the multimodal representation. After that, it applies a tensor decomposition approach by calculating the inner product of the multimodal tensor with a weight tensor. The output is a low-dimension vector, which is used to make predictions.

Seq-to-Seq: Multimodal Transformer (MulT) [126] merges multimodal time-series via a feed-forward fusion process from multiple directional pairwise cross-modal transformers. Each cross-modal transformer is a deep stacking of several cross-modal attention blocks. As a final step, it concatenates the outputs from the cross-modal transformers and passes the multimodal representation through a sequence model to make predictions.

Quantum-inspired: QMF [73] is the model introduced in chapter 4.

Experimental Settings

A grid search for the best hyper-parameters was conducted for all models. At each search, the models were trained for 100 epochs. Out of 50 searches, the model with the lowest validation loss was used to produce the test performance. The parameters in the proposed EFNN model were determined by the set of hyper-parameters $\Theta = \{D, K\}$, where D is the embedding dimension of input features into same dimensional spaces and K is the number of measurement vectors. For both datasets, we searched over a parameter pool : $D \in \{100, 120, 140, 160, 180, 200\}$, $K \in \{10, 20, 30, 60, 80, 120, 150\}$, size of last hidden layer $\in \{32, 48, 64, 80\}$, dropout rate for the last hidden layer $\in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$, learning rate $\in \{0.001, 0.002, 0.005, 0.008\}$, batch size$\in \{32, 64, 96\}$.

We trained EFNN by feeding the real and imaginary parts of the complex-valued layers as different input parts and simulated complex operations using real values [125]. EFNN was hence trained via the backpropagation algorithm. Measurements were initialized from standard normal distributions. All the parameters were trainable with respect to $L1$-loss defined on the extracted features. We chose Adam as the optimization algorithm.

Performance Analysis

Table 5.8 shows the comparison results between EFNN and the SOTA baseline approaches for the CMU-MOSI task. The approaches that apply attention mechanism to align pairwise modalities, i.e., c-GRU+Att and MulT, exhibit the highest binary accuracy as compared to the rest of the baselines. TFN achieves the highest accuracy for Acc_7 among the baselines. Note that there is a discrepancy between the empirical results from our experiments and the reported ones in the literature. Specifically, we empirically find a lower accuracy for all the SOTA approaches. A possible reason for the discrepancy between literature and our empirical results may be that different versions of the CMU-MOSI dataset had been used in the published works. Those versions consist of different feature dimensions and sequence lengths. Another possible explanation for this might be the fine-tuning parameters, which are rarely reported in the existing work, making reproducibility a particularly tricky task.

Table 5.9 presents the results for the CMU-MOSEI task. All approaches attain an improved performance compared to that of the CMU-MOSI dataset. We suspect this is because CMU-MOSEI is a much larger dataset. c-GRU+Att is the most effective model among the baselines for the CMU-MOSEI task. MulT achieves similar performance to c-GRU+Att, without a significant difference. EFNN gains an increased binary accuracy of 82.8% as compared to 80.7% of c-GRU+Att, which is a significant improvement of 2.6% (t-test$<$.05). Finally, EFNN achieves an improvement for all evaluation metrics on CMU-MOSEI.

In summary, EFNN significantly outperforms the baselines for both CMU-MOSI and CMU-MOSEI tasks. The analysis of results has shown that EFNN is capable of coping with both balanced and skewed datasets.

Approach	Acc_7	Acc_2	$F1$	MAE	$Corr$
Baseline LSTM					
EF-LSTM	32.7	75.8	75.6	1.000	0.630
LF-LSTM	32.7	76.2	76.2	0.987	0.624
Advanced LSTM					
MARN [153]	31.8	76.4	76.2	0.984	0.625
MFN [152]	31.9	76.2	75.8	0.988	0.622
c-GRU+Att [47]	33.8	78.2	78.1	0.947	0.675
Tensor					
TFN [151]	34.9	75.6	75.5	1.009	0.605
LMF [85]	30.5	75.3	75.2	1.018	0.605
Seq-to-Seq					
MulT [126]	33.6	78.7	78.4	0.964	0.662
Quantum					
QMF [73]	34.2	78.1	77.9	0.997	0.670
EFNN	**35.9**	**80.9**	**80.8**	**0.913**	**0.690**
(Δ%)	(2.8%)	(2.7%)	(2.9%)	(3.7%)	(2.2%)
($\Delta_{EF-LSTM}$%)	(8.9%)	(6.3%)	(6.4%)	(9.5%)	(8.7%)

Table 5.8: Effectiveness on CMU-MOSI. Best results are highlighted in bold. (Δ%) and ($\Delta_{EF-LSTM}$%) indicate absolute relative percentage improvement over the next best model and the baseline EF-LSTM model, respectively.

Ablation Test

We also carried out an ablation test on CMU-MOSEI to investigate the effect of introduced quantum components. In particular, to examine the effectiveness of convolution neural networks CNN_m projecting modalities to common dimensional spaces, we replace the component with GRU layers (a.k.a. *EFNN-gru*). Furthermore, we would like to investigate the impact of non-separable modalities, by introducing two other variants of EFNN, after removing the weight vector W (see Figure 5.8): a) *EFNN-tensor* fuses all modalities into a unified tensor-based representation, i.e., tri-modal fusion; and b) *EFNN-con* concatenates all modalities into a vector representation, and then the outputs interact with the measurement component. Moreover, we also consider the impact words' position in an utterance by initializing phases from standard normal distributions (a.k.a. *EFNN-rand*). We finally replace the measurements with a convolutional neural network (CNN), whereby the K filters of CNN serve as K measurements, in order to investigate the impact of the measurement component (a.k.a. *EFNN-cnn*).

The results of the ablation test, illustrated in Table 5.10, show that each component plays a crucial role in the EFNN. In particular, the comparison with EFNN-rand shows the effectiveness of modelling words' position into the phase part of complex-valued representations. At the same time, the decreased performance of EFNN-tensor and EFNN-con reveals the superiority of encoding the non-classical correlations (i.e., entanglements) between modalities. Moreover, the comparison

Approach	Acc_7	Acc_2	$F1$	MAE	$Corr$
Baseline LSTM					
EF-LSTM	45.7	78.2	77.1	0.687	0.573
LF-LSTM	47.1	79.2	78.5	0.655	0.614
Advanced LSTM					
MARN [153]	47.7	79.3	77.8	0.646	0.629
MFN [152]	47.4	79.9	79.1	0.646	0.626
c-GRU+Att [47]	48.4	80.7	80.2	0.627	0.672
Tensor					
TFN [151]	47.3	79.3	78.2	0.657	0.618
LMF [85]	47.6	78.2	77.6	0.660	0.623
Seq-to-Seq					
MulT [126]	46.6	80.2	79.8	0.636	0.654
Quantum					
QMF [73]	47.2	79.8	79.4	0.646	0.655
EFNN	**50.2**	**82.8**	**82.6**	**0.595**	**0.689**
(Δ%)	(3.6%)	(2.6%)	(2.9%)	(5.4%)	(2.5%)
($\Delta_{EF-LSTM}$%)	(9.0%)	(5.6%)	(6.7%)	(15.5%)	(16.8%)

Table 5.9: Effectiveness on CMU-MOSEI. Best results are highlighted in bold. (Δ%) and ($\Delta_{EF-LSTM}$%) indicate absolute relative percentage improvement over the next best model and the baseline EF-LSTM model, respectively.

Approach	Acc_7	Acc_2	$F1$	MAE	$Corr$
EFNN-gru	1.5% (↓)	1.3% (↓)	1.3% (↓)	1.2% (↓)	1.1% (↓)
EFNN-tensor	2.2% (↓)	1.8% (↓)	1.8% (↓)	1.5% (↓)	1.7% (↓)
EFNN-con	2.8% (↓)	2.5% (↓)	2.5% (↓)	2.4% (↓)	2.2% (↓)
EFNN-rand	1.5% (↓)	1.2% (↓)	1.2% (↓)	1.2% (↓)	1.4% (↓)
EFNN-cnn	1.9% (↓)	1.4% (↓)	1.4% (↓)	1.6% (↓)	1.7% (↓)

Table 5.10: Ablation test on CMU-MOSEI. Absolute relative percentage difference from EFNN.

with EFNN-cnn shows the usefulness of trainable measurements. Finally, EFNN-gru shows that convolutional neural networks could be a more appropriate way to project modalities into common dimensional spaces. Overall, the ablation test reveals that the entanglement-driven fusion component plays the most crucial role in the architecture of EFNN.

Post-hoc Interpretability

Further, we evaluated the post-hoc interpretability by investigating the bi-modal correlations within composite utterance states after the modality context interaction. In particular, according to Equation 3.11, we calculated the degree of quantum entanglement for bipartite composite utterance states of linguistic and visual modalities.

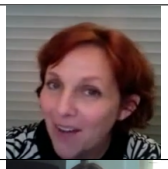
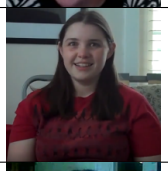
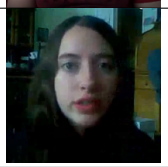
Linguistic	Visual	Sentiment
The story was all right.		Positive
But it does have some adult humour.		Positive
I do not wanna see any more of this.		Negative

Table 5.11: Selected most entangled linguistic-visual modalities on CMU-MOSI.

Table 5.11 and Table 5.12 illustrate some examples of the most and least entangled linguistic-visual modalities according to entanglement entropy. The most entangled pairs are those that one of two modalities is ambiguous or uninformative. For example, in Table 5.11, the linguistic content of the two first utterances is ambiguous, while the visual content of the third utterance is uninformative. By contrast, when the context of both modalities is informative, unambiguous, and simultaneously present (see Table 5.12), the entanglement entropy is close to zero. In those cases, the composite representation is separable, and there is no need to exploit the quantum probabilistic interpretation. However, through the concept of non-separability, EFNN is able to capture both separable and non-separable bi-modal interactions, as a generalization of existing probabilistic modality fusion approaches. This attribute is the core reason that EFNN has achieved performance improvement.

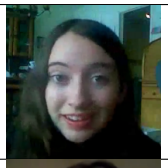
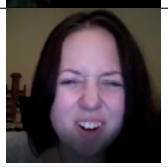
Linguistic	Visual	Sentiment
The voice acting was phenomenal!		Positive
Yeap a horrible protagonist!		Negative

Table 5.12: Selected examples of least entangled linguistic-visual modalities on CMU-MOSI.

5.3.4 Section Conclusions

We have proposed a variance of c-EFNN, namely EFNN, discarding the context modelling component, which considers preceding utterances. Comprehensive experiments on two SOTA benchmarking datasets showed that the information encoded in the form of non-classical correlations between modalities is not only effective for the utterance-level modality fusion but also for the word-level modality fusion. Moreover, the concept of non-separability provides additional cues about the interactions across different modalities.

5.4 Chapter Conclusions

In this chapter, we have presented an investigation of modelling cross-modal information in the form of non-classical correlations. We have initially introduced a methodology for modelling pairwise decisions of documents and then investigating the existence of non-classical correlations via the violation of CHSH inequality to facilitate the decision-level modality fusion. However, the experiments on an image-text dataset demonstrated that there are no such correlations between pairs of documents. That is mainly due to the lack of interactions between documents and complex-field. Motivated by these results, we have developed an end-to-end and transparent quantum probabilistic fusion neural network for video emotion recognition and sentiment analysis. Comprehensive experiments on SOTA benchmarking datasets have shown the effectiveness of modelling cross-modal information in the form of non-classical correlations for both word-level and utterance-level fusion tasks, as well. Moreover, non-classical correlations have been quantified by appropriate measures, which optimized post-hoc interpretability. Despite the encouraging results, it remains an open research question of how multimodal decision perspectives could be modelled and learnt in a quantum manner to leverage the expensiveness of quantum probabilities. In the next chapter, we go one step forward and answer this question by introducing a novel quantum decision-level representation and proposing a methodology to learn such representation from multimodal data.

Chapter 6

Quantum-inspired Decision-level Multimodal Fusion

In Section 5.1, we have proposed a methodology to model decisions of multimodal documents with a quantum view to investigate the existence of non-classical correlations via the violation of CHSH inequality. However, due to the lack of interactions and complex field, no violation has been observed. Additionally, the analysis in Chapter 2 showed that video sentiment analysis as a decision-making process is inherently complex, involving the fusion of decisions from multiple modalities and human cognitive biases. Inspired by recent advances in the emerging field of quantum cognition [41, 131], we show that the sentiment judgment from one modality could be incompatible with the judgment from another, i.e., the order matters and they cannot be jointly measured to produce a final decision. Thus the cognitive process in video sentiment analysis exhibits "quantum-like" biases that cannot be captured by classic probability theories. Accordingly, in Chapter 6, we propose a fundamentally new, quantum cognitively-motivated fusion strategy for predicting sentiment judgments. In particular, we have formulated utterances as quantum superposition states of positive and negative sentiment judgments, and uni-modal classifiers as mutually incompatible observables, on a complex-valued Hilbert space with positive-operator valued measures. Experiments on two benchmarking datasets illustrate that our proposed model significantly outperforms various existing decision level and a range of SOTA content-level fusion approaches. The results also show that the concept of incompatibility allows effective handling of all combination patterns, including those extreme cases that are wrongly predicted by all uni-modal classifiers.

6.1 Introduction

Video sentiment analysis is an emerging interdisciplinary area, bringing together artificial intelligence (AI) and cognitive science. It studies a speaker's sentiment expressed by distinct modalities, i.e., linguistic, visual, and acoustic. At its core, effective modality fusion strategies are in place. Existing neural structures achieve the SOTA [38, 126, 143, 151] by integrating features after being extracted, called *content-level* fusion. Other approaches simulate logic reasoning and human cognitive biases [54, 98] by aggregating decisions of uni-modal classifiers into a joint decision, called *decision-level* fusion. Additionally, hybrid fusion approaches benefit from the advantages of both strategies. In this work, we target at the generally less effective but more flexible decision-level fusion.

Video sentiment analysis is inherently a complex human cognition process. Recent research in cognitive science found that in some cases human decision making could be highly irrational [129], and such behaviour does not always obey the classical (Kolmogorov) axioms of probability [68] and utility theory [97]. On the other hand, preliminary work shows that the mathematical formalism of Quantum Mechanics (QM) can successfully address paradoxes of classical probability theory in modelling human cognition [25]. Particularly, in [140], the Quantum Question Order inequality was tested, which is an inequality for testing incompatibility, and hence irrational behaviour in decision-making systems. In this study, users were asked to judge the relevance of documents. The results of the study showed that the relevance was affected by the order in which documents were presented. For example, for the query "Albert Einstein", users were shown documents about "Isaac Newton" and "Theory of Relativity". The relevance probability of "Isaac Newton" was lower when it was shown after "Theory of Relativity" (called a comparative context) than when it was shown first (non-comparative context). In simple terms, having seen a more relevant document first, user's judgement about a particular document may change. This can be explained as an Order Effect due to incompatibility between the topics. Conceptually, quantum cognition challenges the notion that user's cognitive states underpinning the decisions have pre-defined values and that a measurement merely records them. Instead, the cognitive system is fundamentally uncertain and in an indefinite state. The act of measurement would then create a definite state out of the indefinite state and change the overall cognitive state.

Likewise, we hypothesise that uni-modal sentiment judgments do not happen independently, like a pre-defined value being read out of the internal cognitive state. They are rather constructed at the point of information interaction and thus influenced by the other modalities, which serve as a context for the inference of sentiment judgment for the current modality. For example, there might be cases that the order of different decision perspectives, e.g., when someone focuses first on the linguistic and then on the visual perspective, or vice versa, can lead to controversial sentiment judgements. That is, the measurement from the first perspective provides a context that affects the subsequent one, influencing the probabilities used to compute the utility function of multimodal sentiment decision. In this case, we say that these two decision perspectives are *incompatible* with each other.

That implies that judgements over different modalities cannot be measured jointly, and quantum probability should hence be in place [133]. We argue that video sentiment analysis could benefit from the generalized framework of quantum cognition by capturing the cases of incompatibility which cannot be modelled by classical probabilities.

To this end, we introduce a novel decision-level fusion model inspired by quantum cognition [41]. The goal is to predict the sentiment of utterances in videos, associated with linguistic, visual, and acoustic streams. In particular, we formulate an utterance as a *quantum superposition* state of positive and negative sentiments (i.e., it can be positive and negative at the same time until it is judged under a specific context), and uni-modal classifiers as *mutually incompatible observables*, on a shared complex-valued Hilbert space $\mathcal{H}$ spanned by distinct uni-modal sentiment bases. In QT, the concept of incompatibility of observables, represented in the form of non-commuting operators, implies that it is impossible to construct a joint probability distribution for the variables. We can only assign probabilities to a sequence of measurements. That is, a quantum state is impossible to be in a definite state with respect to three incompatible bases because the definite state for a specific basis results in an indefinite state for the other bases. In this chapter, we take advantage of incompatibility to influence the uni-modal decisions, when they are under high uncertainty, to finally infer multimodal sentiment judgments. To resolve the incompatibility issue, we make use of Positive-Operator Valued Measures (POVMs) to approximate the sentiment of uni-modal classifiers simultaneously. In practice, we estimate the complex Hilbert Space and uni-modal observables from training data, and then establish the final multimodal sentiment state of a test utterance from the learned uni-modal observables. It is important to note that the model produces a generalization form of classical probabilities, allowing for both compatible and incompatible sentiment decisions.

To our best knowledge, this is the first quantum cognitively inspired theoretical approach, with practical implementation, that investigates and models the incompatibility of sentiment judgments for video sentiment analysis.

Extensive evaluation on two widely used benchmarking datasets, namely CMU-MOSI[155] and CMU-MOSEI[157], show that our model significantly outperforms not only various representative decision-level fusion baselines, but also a range of SOTA content-level fusion approaches for video sentiment analysis. We also show that the model is able to make correct sentiment judgments even for the cases where all uni-modal classifiers give wrong predictions.

6.2 Background

In chapter 3, we introduced the fundamental concepts of QT, such as *Hilbert Space*, *Quantum Superposition*, and *Quantum Measurement* that we have exploited to construct the proposed model.

In contrast to previous quantum models introduced in chapters 4 and 5, the current quantum-probabilistic model takes advantage of quantum incompatibility to fuse uni-modal sentiment judgments. To that end, the next part introduces the concept of incompatibility.

6.2.1 Incompatibility

The concept of *incompatibility* is applicable in a Hilbert space only. Each basis state, defining an event, has a projector Π to evaluate the event. In contrast to classical probability, the conjunction of two events is not necessarily commutative [26]. Suppose Π_A and Π_B are two sequential measurements for A and B events. In quantum cognition, the joint probability distribution of two events equals the product of the two projectors Π_A and Π_B, corresponding to the basis state $A \cap B$. If $\Pi_A\Pi_B = \Pi_B\Pi_A$, then the two events are called *compatible*. However, if $\Pi_A\Pi_B \neq \Pi_B\Pi_A$, then their product is not a projector, and the two events do not commute, that is, they are *incompatible*. Incompatibility implies that the two measurements cannot be accessed jointly without disturbing each other. Classical probability can not capture such disturbance, assuming that measurements are always compatible and commute. However, the mathematical formalism of quantum probability allows for both compatible and incompatible measurements [64, 133]. Thus, it is a generalization of classical probability theory.

6.3 Model

The proposed model draws an analogy from quantum systems. For instance, consider an electron and suppose we want to measure a particular property, called the spin. The spin is a magnetic property. We could hence attribute a positive spin if the electron deflects towards the North pole and negative if it deflects towards the opposite pole. However, the electron is a physical entity described by other properties as well, such as momentum and energy. Likewise, we could attribute positive and negative momentum and energy. Crucially, positive and negative values of a property, e.g., spin, are not independent of the other properties, e.g., momentum and energy. The choice of basis that is made in formulating a property is what leads to the expectation that QT might be used in the description of contextual systems. To draw the analogy of the electron property states in terms of human multimodal sentiment judgments, we considered the two-valued property data to be equivalent to the positive/negative sentiment judgments. Moreover, the different properties represented by bases are equivalent to making judgments along distinct modalities, i.e., linguistic, visual, and acoustic. Just like the example of the electron, the final multimodal sentiment judgment cannot be assumed to exist independently of the choice of a specific modality considered. In the remaining part of the section, we elaborate on the procedural steps of the introduced model. The task is an utterance-level sentiment analysis, as discussed in Chapter 2, Section 2.2.1.

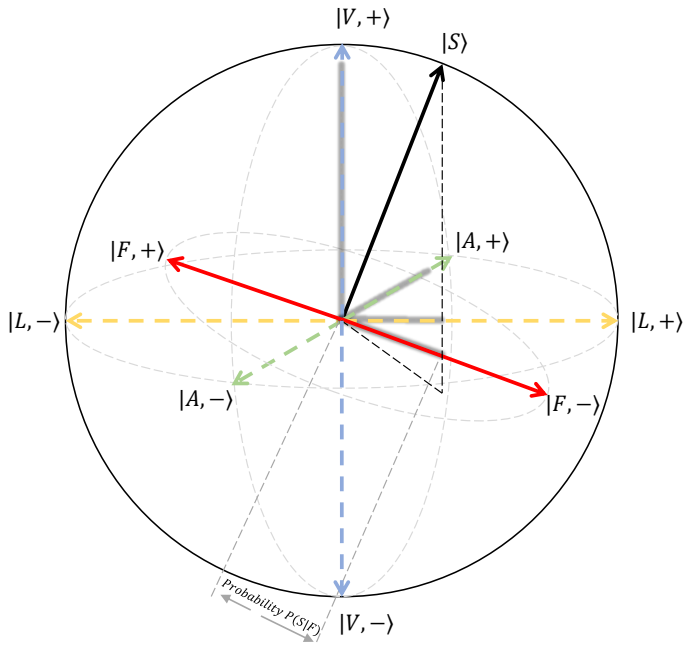


Figure 6.1: The Sentimental Hilbert Space. An utterance is represented as a pure state $|S\rangle$ belonging to the surface of a unit sphere (called the Bloch sphere). The two opposed unit vectors represent positive and sentiment judgment, and the ellipses represent subspaces, i.e., events. The associated uni-modal sentiment observables $\hat{L}, \hat{V}, \hat{A}$ as well as the tri-modal observable $\hat{F}$ are mutually incompatible in that they have different eigenstates. The observables are not orthogonal since modalities are not independent but highly correlated. Shadowed basis vectors imply projections of $|S\rangle$ on the corresponding bases, i.e., probability of events.

6.3.1 Sentiment Hilbert Space

The model is defined on a *Sentimental Hilbert Space* $\mathcal{H}_{senti}$, which is a 2-dimensional vector space spanned by the basis states $\{|+\rangle, |-\rangle\}$. The basis states $|+\rangle$, $|-\rangle$ correspond to positive (if the human annotation score ≥ 0) and negative (if the human annotation score < 0) sentiments, respectively. We represent an utterance U_k as a pure state $|S_{U_k}\rangle$ (in short $|S\rangle$) on $\mathcal{H}_{senti}$. The uni-modal sentiment classifiers (denoted as L, V, A) are formulated as mutually incompatible observables (see Figure 6.1). The utterance can be represented under different sets of basis states, i.e., uni-modal (L, V, A) and multimodal (F) basis states in Figure 6.1. The observables are not orthogonal since modalities are not independent, but highly correlated. For tasks beyond binary sentiment classification, i.e., emotion recognition, we follow a one-vs-all strategy for each emotion by recognizing one emotion from all the others.

6.3.2 Utterance Representation

An utterance is represented as a pure state $|S\rangle$ of positive and negative sentiments on $\mathcal{H}_{senti}$ (see Figure 6.1). On this 2-dimensional Hilbert space, it adopts a polar generic representation

$$|S\rangle = \cos\frac{\theta_S}{2}|+\rangle + e^{i\phi_S}\sin\frac{\theta_S}{2}|-\rangle, \tag{6.1}$$

where $\theta_S, \phi_S \in [0, 2\pi]$ and and i is the imaginary number satisfying $i^2 = -1$. According to the Born's rule [64], the probability of an utterance being positive and negative is $P(+) = |\langle S|+\rangle|^2 = \cos^2\frac{\theta_S}{2}$ and $P(-) = |\langle S|-\rangle|^2 = \sin^2\frac{\theta_S}{2}$, with $\cos^2\frac{\theta_S}{2} + \sin^2\frac{\theta_S}{2} = 1$. As to be shown in more detail in the next Section, the *relative phase* ϕ_S plays a crucial role in capturing correlations between incompatible observables and giving rise to results that are fundamentally different from the classical case.

6.3.3 Sentiment Decisions

We formulate uni-modal sentiment decisions as mutually incompatible observables on $\mathcal{H}_{senti}$, namely $\hat{L}$, $\hat{V}$, and $\hat{A}$ for linguistic, visual and acoustic modalities respectively (see Figure 6.1). For the binary sentiment analysis task, each observable is associated with two eigenstates and two eigenvalues, with common eigenvalues of 1 and -1 for positive and negative sentiments. In that case, incompatibility falls under different sets of eigenstates $\{|M,+\rangle, |M,-\rangle\}$ defining a uni-modal basis, where modality $M \in \{L, V, A\}$. Following Equation 6.1, we express the eigenstates as

$$|M,+\rangle = \cos\frac{\theta_M}{2}|+\rangle + e^{i\phi_M}\sin\frac{\theta_M}{2}|-\rangle \tag{6.2}$$

$$|M,-\rangle = \sin\frac{\theta_M}{2}|+\rangle - e^{i\phi_M}\cos\frac{\theta_M}{2}|-\rangle \tag{6.3}$$

with $\theta_M, \phi_M \in [0, 2\pi]$. The eigenstates form an orthonormal basis, with $\langle M,+|M,+\rangle = \langle M,-|M,-\rangle = 1$ and $\langle M,+|M,-\rangle = \langle M,-|M,+\rangle = 0$.

A general observable $\hat{O}$ can be decomposed to its eigenstates $\{|\lambda_i\rangle\}$ of the orthonormal basis as $\hat{O} = \lambda_i |\lambda_i\rangle\langle\lambda_i|$, where eigenvalues $\{\lambda_i\}$ are possible values that a state can take for the corresponding events after measurement. Thus, the uni-modal observables are defined as follows:

$$\hat{L} = (+1)|L,+\rangle\langle L,+| + (-1)|L,-\rangle\langle L,-| \tag{6.4}$$

$$\hat{V} = (+1)|V,+\rangle\langle V,+| + (-1)|V,-\rangle\langle V,-| \tag{6.5}$$

$$\hat{A} = (+1)|A,+\rangle\langle A,+| + (-1)|A,-\rangle\langle A,-| \tag{6.6}$$

Similarly, the observable for the final sentiment decision $\hat{F}$ is

$$\hat{F} = (+1)|+\rangle\langle+| + (-1)|-\rangle\langle-| \tag{6.7}$$

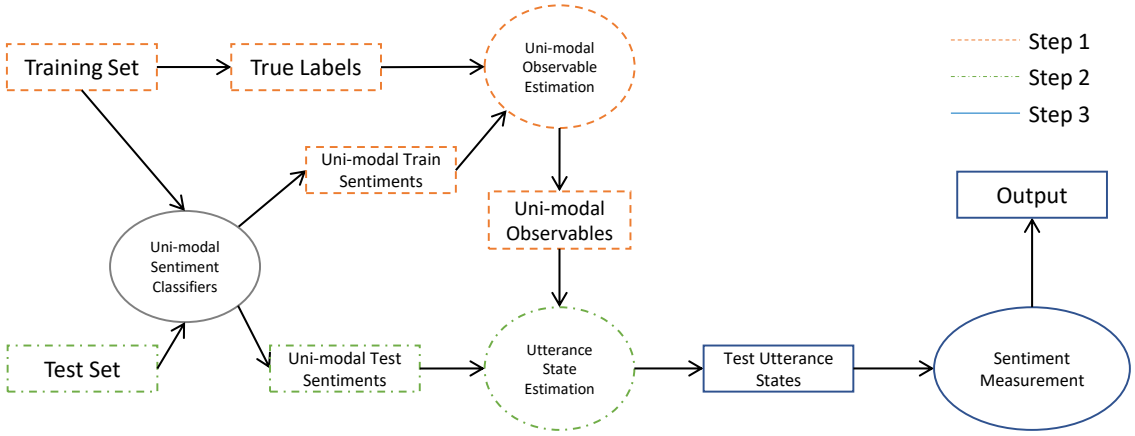


Figure 6.2: Model pipeline. It consists of three steps: a) step 1: Observable estimation, b) step 2: Utterance state estimation, and c) step 3: Multimodal sentiment measurement.

which spans the $\mathcal{H}_{senti}$ and is incompatible with all uni-modal observables.

Following the projective geometric structure, the result probability of a measurement on an eigenstate equals the projection of the state onto it, that is, the squared inner product of the vectors, e.g., $|\langle S|M,+\rangle|^2$ for uni-modal positive sentiment and $|\langle S|+\rangle|^2$ for final (multimodal) positive sentiment. The measurement probabilities under $\hat{L}$ stand for the utterance's sentiment under linguistic modality, as so forth for the other modalities. Finally, its multimodal sentiment is determined by the observable $\hat{F}$. As illustrated in Figure 6.1, the sentiment judgment is positive in terms of uni-modal observables (projections are visualized as shadows in Figure 6.1), yet is negative in terms of the multimodal observable due to incompatibility.

6.4 Methodology

This section presents a methodology that operationalizes the proposed model. Traditionally, in the physical sciences, the study of mathematical problems involves modelling methods leveraging a combination of approximation techniques. In this work, we exploit statistical information from the data to learn the sentimental Hilbert Space described in the previous section, so as to leverage the incompatible observables to determine the sentiment of an utterance. Overall, as shown in Figure 6.2, the pipeline consists of three steps: (1) we first estimate the generic uni-modal observables $\hat{M}$ from the training data; (2) then we construct the sentiment state for each test utterance $|S_T\rangle$ from the learned uni-modal observables and uni-modal sentiment prediction results; (3) finally, we judge the sentiment from the multimodal observable $\hat{F}$. In the remaining part of the section, we elaborate the methodology for each step.

6.4.1 Observable Estimation

The uni-modal observables are constructed from the overall statistics of the training data. These values are mapped to their quantum expressions to estimate the parameters of the uni-modal observables. In particular, the uni-modal observables and pure state should submit to the following properties: I) the pure state should conform to the statistics of the dataset, II) the uni-modal sentiment measurement results should be consistent with the ratio of positive and negative samples in the training subsets, and III) quantum correlations between observables should be aligned to classical correlations of the per-sample prediction results, derived from the training data.

To facilitate the construction of uni-modal observables, we introduce a pure state as follows:

$$|G\rangle = \cos\frac{\theta_G}{2}|+\rangle + e^{i\phi_G}\sin\frac{\theta_G}{2}|-\rangle, \tag{6.8}$$

which describes the extent to which the dataset is unbalanced for positive and negative labels. By Born's rule [64], the probability of positive judgment is:

$$P(+) = |\langle +|G\rangle|^2 \approx \frac{\#pos}{N}, \tag{6.9}$$

where $\#pos$ is the number of true positive utterances in the training set and N the size of the training set. Equation 6.9 implies

$$\cos^2\frac{\theta_G}{2} \approx \frac{\#pos}{N}, \tag{6.10}$$

since the quantum probability equals the squared amplitude of a state (see chapter 3).

According to the second property, the probability of a positive sentiment judgment for each modality is given by

$$P_M(+) = |\langle M,+|G\rangle|^2 \approx \frac{\#M_{pos}}{N}, \tag{6.11}$$

where $\#L_{pos}$, $\#V_{pos}$ and $\#A_{pos}$ equals the number of true positive utterances for each modality in the training set. Combining Equation 6.2, Equation 6.8 and Equation 6.11, the probability of the positive sentiment judgment for each modality is

$$\cos^2\frac{\theta_M}{2}\cos^2\frac{\theta_G}{2} + \sin^2\frac{\theta_M}{2}\sin^2\frac{\theta_G}{2} + \frac{1}{2}\sin\theta_M\sin\theta_G\cos(\phi_M-\phi_G) \approx \frac{\#m_{pos}}{N}. \tag{6.12}$$

Finally, we looked into the correlations between pairs of uni-modal observables, where the relative phases play an important role. From a quantum statistics point of view, the correlation of observables for two modalities M_1, M_2 is given by $(|\langle M_1,+|M_2,+\rangle|^2 + |\langle M_1,-|M_2,-\rangle|^2 - |\langle M_1,+|M_2,-\rangle|^2 - |\langle M_1,-|M_2,+\rangle|^2)$. A more detailed explanation about the derivation is included in Appendix C.

It should in principle be aligned to the classical correlations derived from the data. Hence we have

$$\frac{1}{2}\big(|\langle M_1,+|M_2,+\rangle|^2+|\langle M_1,-|M_2,-\rangle|^2-|\langle M_1,+|M_2,-\rangle|^2-|\langle M_1,-|M_2,+\rangle|^2\big)\approx corr(M_1,M_2), \tag{6.13}$$

where $M_1 \neq M_2 \in \{L, V, A\}$ and $corr(M_1, M_2)$ is a classical correlation of the per-sample prediction results based on modalities M_1 and M_2, which is computed from the training data. When M_1 and M_2 give exactly same predictions, the correlation $corr(M_1, M_2) = 1$. Therefore, $|\langle M_1,+|M_2,+\rangle| = |\langle M_1,-|M_2,-\rangle|^2 = 1$ and $|\langle M_1,+|M_2,-\rangle|^2 = |\langle M_1,-|M_2,+\rangle|^2 = 0$, so the value 1 is also produced from the quantum side. Similarly, a value of -1 is obtained for both sides when the two modalities give totally opposite predictions, indicating the maximum negative correlation. Hence, Equation 6.13 gives three equations for distinct pairs of modalities. For example, for linguistic-visual correlation, Equation 6.13 results in

$$\cos\theta_L\cos\theta_V+\sin\theta_L\sin\theta_V\cos(\phi_L-\phi_V)\approx corr(L,V), \tag{6.14}$$

as so forth for the $\{L, A\}, \{A, V\}$ modality pairs respectively, that is,

$$\begin{aligned}\cos\theta_L\cos\theta_A+\sin\theta_L\sin\theta_A\cos(\phi_L-\phi_A)&\approx corr(L,A)\\ \cos\theta_A\cos\theta_V+\sin\theta_A\sin\theta_V\cos(\phi_A-\phi_V)&\approx corr(A,V)\end{aligned} \tag{6.15}$$

To wrap up, taking into account the number of positive sentiments in the training set and correlations across different pairs of modalities, we get seven equations from Equation 6.10, Equation 6.12 and Equation 6.13, and eight unknown variables $\{\theta_G, \theta_L, \theta_V, \theta_A, \phi_G, \phi_L, \phi_V, \phi_A\}$. As all equations rely only on the differences between the relative phases rather than their absolute values, we can safely set $\phi_G = 0$ without loss of information. Hence a unique solution of $\{\theta_G, \theta_L, \theta_V, \theta_A, \phi_L, \phi_V, \phi_A\}$ can be produced, and accordingly determining the uni-modal observables $\hat{L}$, $\hat{V}$, and $\hat{A}$.

6.4.2 Utterance State Estimation

After having uni-modal observables calculated as described above, we need to estimate the state for each test utterance. For a specific test utterance denoted as

$$|S_T\rangle=\cos\frac{\theta_T}{2}|+\rangle+e^{i\phi_T}\sin\frac{\theta_T}{2}|-\rangle, \tag{6.16}$$

the uni-modal predictions can be exploited to estimate the values of θ_T, ϕ_T. However, since the observables $\hat{L}$, $\hat{V}$, and $\hat{A}$ are mutually incompatible, the measurements results cannot be accessed simultaneously. To that end, we propose utilize POVMs to get the results of all incompatible measurements simultaneously [130]. In particular, we construct sample-specific POVMs for each uni-modal measurement, applying unsharp (weak) projections [24] without disturbing the observables.

That is, the measurement of each utterance state with respect to a specific-modality basis is modelled as POVM. The POVM on a specific-modality basis does not disturb the measurements on the other two specific-modality bases. Finally, the operators are constructed as follows:

$$E_{\pm}^{M} = \frac{\eta_T}{2}\mathbb{I} + (1-\eta_T)\left|M,\pm\right\rangle\left\langle M,\pm\right|, \tag{6.17}$$

where $\eta_T \in [0,1]$ is specific to sample T, since each utterance interacts with the apparatus in a different manner. We apply uni-modal POVMs on the test utterance to measure the sentiment of utterance in terms of each modality, that is,

$$\left\langle S_T\right| E_{+}^{M} \left|S_T\right\rangle \approx P_{T,M}(+), \tag{6.18}$$

where $P_{T,M}(+)$ are uni-modal probabilities for the positive sentiment judgment. Equation 6.18 gives a system with three equations, each equation for a distinct modality, and three unknown variables $\{\theta_T, \phi_T, \eta_T\}$. Solving the system allows us to construct the state $|S_T\rangle$.

6.4.3 Multimodal Sentiment Measurement

The sentiment of a test utterance $|S_T\rangle$ is measured by Equation 6.7. The results are $P_T(+) = \cos^2\frac{\theta_T}{2}$ and $P_T(-) = \sin^2\frac{\theta_T}{2}$. The sentiment of S_T is hence positive if $\cos^2\frac{\theta_T}{2} > 0.5$ and negative otherwise.

6.5 Experiments

6.5.1 Datasets

We evaluate the proposed model on two affective analysis tasks. For video sentiment analysis, we have performed experiments on two benchmarking datasets, namely, CMU Multimodal Opinion-level Sentiment Intensity (CMU-MOSI) [155] and CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) [157]. We have left the details about the dataset and feature extraction in section 4.4.1. For video emotion recognition, we conducted experiments on IEMOCAP [27]. We have left more details about IEMOCAP and feature extraction in section 2.2.2.

For evaluation metrics, we have chosen the most stabilized metrics from the full set given in Section 2.2.3. For example, the experiments in Chapter 2 showed that Acc_7 and $Corr$ metrics are not stabilized enough as compared to Acc_2 and $F1$ metrics. Additionally, in contrast to Acc_7, the Acc_2 facilitates the training process, since for evaluating Acc_7, we need to train the model for each of the 7 classes, following a one-vs-all strategy. In particular, we have adopted the binary accuracy (i.e., Acc_2 : positive sentiment if the human annotation score ≥ 0, and negative sentiment if the human annotation score < 0), and $F1$ score for CMU-MOSI and CMU-MOSEI tasks. For IEMOCAP, we have followed a one-vs-all strategy. Thus, we have again exploited the binary accuracy and F_1 score

as evaluation metrics for IEMOCAP task.

6.5.2 Baselines

We compared with robust approaches on both decision-level and content-level modality fusion approaches.

Decision-level:

We first trained neural uni-modal classifiers. In particular, we used Bi-GRU layers [33] with forward and backward state concatenation, followed by fully connected layers. The outputs gave linguistic, visual, and acoustic embeddings $\{L, V, A\} \in \mathbf{R}^d$, where d was the number of neurons in dense layers. Then, self-attentions were computed for each uni-modal dense representation, by calculating the scaled dot-product [135]. Finally, each attentive uni-modal representation was fed into two fully connected layers, followed by a softmax layer to obtain sentiment judgments. We leave the final settings in Appendix C.1. The uni-modal results are then fed into the multimodal meta-fusion approaches. We compared with a range of baseline fusion approaches:

- **Voting** was used to aggregate the outputs of the uni-modal classifiers. In particular, we applied a) *Hard Voting*, via majority voting, b) *Weighted Majority Voting*, by assigning weights to each uni-modal classifier and taking their average, and c) *Soft Voting*, by averaging the predicted probabilities, to infer multi-modal sentiment judgments.

- **Single models** exploited supervised machine learning algorithms as meta fusion approaches of the uni-modal classifiers. For both tasks, we chose the most effective models, namely, a) *Logistic Regression*, b) *Support Vector Machine (SVM)*, and c) *Gaussian Naive Bayes (GaussianNB)*, from a pool of supervised learning algorithms.

- **Ensemble methods** combined learning algorithms, selecting the optimum combination from a pool of models. We explored stacking, backing, and boosting strategies [111]; a) for stacking, single models were stacked together and the hard voted method computed predictions, b) for bagging, a number of estimators were aggregated by majority voting, and c) for boosting, we applied AdaBoost classifier [42].

- **A Deep Fusion** approach combined the confidence scores of uni-modal classifiers along with the complementary scores as inputs to a deep neural network, followed by a sigmoid layer, which made the final prediction [101].

Content-level:

We also compared the model with a range of SOTA content-level fusion approaches.

- For **SOTA,** we replicated a) *MulT* [126], consisting of pairwise crossmodal transformers, the outputs of which are concatenated to build the multimodal embedding utterance, b) *RAVEN* [143], an RNN based framework with an attention gating mechanism to model cross-modal interactions, and c) *TFN* [151], a tensor-based neural network that a multi-dimensional tensor captures uni-modal, bi-modal, and tri-modal interactions across distinct modalities.
- **QMF** [73] is the model introduced in chapter 4.

6.5.3 Experiment Settings

We conducted the experiments on the same uni-modal classifiers trained for the decision-level baseline approaches. We estimated the uni-modal observables from training plus validation sets, and then we used the learnt observables for predicting the utterance sentiment on the test set. We used Pearson correlation for modelling classical correlations. In case the equation systems did not have solutions, the MATLAB $fsolve$ function was used to generate a numerical solution. In particular, we randomly initialized the parameters $\{\theta_G, \theta_L, \theta_V, \theta_A, \phi_L, \phi_V, \phi_A\} \in [0, 2\pi]$ for uni-modal observable estimation, and $\{\theta_T, \phi_T\} \in [0, 2\pi], \eta_T \in [0, 1]$ for utterance state estimation. The random initialization was repeated for 200 times to obtain the optimum solutions by calculating the minimum sum of squared loss.

6.5.4 Comparative Analysis of Results

For both video sentiment analysis tasks, we present the comparison results between the proposed model and various decision-level fusion strategies in Table 6.1. For CMU-MOSEI, all approaches attained an improved performance as compared to the performance of CMU-MOSI task. We suspect this is because CMU-MOSEI is a much larger dataset. Overall, Weighted Voting was the best-performing approach among the voting-based aggregations, Logistic Regression among the supervised learning algorithms, and Stacking among the ensemble learning methods. For both tasks, Stacking and Bagging were the most effective baseline decision-level fusion strategies. For CMU-MOSI, the proposed model attained an increased accuracy of 84.6% as compared to 78.4% of Stacking. That is, a significant improvement of 6.2% ($p - value < 0.05$). For CMU-MOSEI, the model reached an increased accuracy of 84.9% as compared to 82.2% of Stacking, i.e., a significant improvement of 2.7% ($p - value < 0.05$). We noticed the $F1$ measure results are pretty different from the binary accuracy in terms of performance. In particular, $F1$ performance is much lower in comparison with Acc_2 performance. We attribute this result to the imbalanced classes in datasets. In particular, further analysis showed that accuracy was very close to precision and quite dissimilar to recall. This means that precision was dominating the overall accuracy. Note that the $F1$ score is the harmonic mean of precision and recall, so it is a class-balanced accuracy measure.

	CMU-MOSI		CMU-MOSEI	
Approach	Acc_2	$F1$	Acc_2	$F1$
Hard Voting	67.5	65.4	71.5	83.3
Weighted Voting	74.6	71.6	81.3	87.8
Soft Voting	75.2	71.9	77.5	86.2
SVM	77.4	72.9	81.7	87.9
Logistic Regression	78.0	73.8	81.9	88.0
GaussianNB	76.7	71.6	80.9	86.8
Stacking	78.4	75.1	82.2	88.1
Bagging	78.1	73.6	82.0	88.0
Boosting	77.7	74.0	81.7	87.7
Deep Fusion	77.8	77.7	81.9	81.3
Proposed Model	**84.6** (↑ 6.2)	**84.5** (↑6.8)	**84.9** (↑ 2.7)	**91.1** (↑3.0)

Table 6.1: Effectiveness of decision-level fusion approaches on CMU-MOSI and CMU-MOSEI. Best results are highlighted in boldface. Numbers in parentheses indicate relative percentage improvement over the next best model.

Table 6.2 presents the comparison results between the introduced model and various content-level fusion approaches. For CMU-MOSI, TFN [151] was the most effective among the baseline content-level fusion approaches. The proposed model attained an improvement in accuracy by 3.4% (see Table 6.2). For CMU-MOSEI, RAVEN [143] attained the highest accuracy among the baselines. The proposed model yielded an increased accuracy of 84.9% as compared to 80.2% of RAVEN, i.e., 4.7% improvement.

Overall, the decision-level feature strategies achieved better performance than the content-level neural approaches on CMU-MOSEI. This implies that discriminative learning approaches can benefit from large datasets, whereas neural approaches lead to overfitting. We also observed that the proposed model achieved a similar performance both on CMU-MOSI and CMU-MOSEI, even though CMU-MOSI is a relatively balanced dataset. That is, our model can effectively cope with both skewed and balanced datasets.

	CMU-MOSI		CMU-MOSEI	
Approach	Acc_2	$F1$	Acc_2	$F1$
MulT [126]	80.2	79.5	80.0	79.8
RAVEN [143]	78.6	78.6	80.2	79.9
TFN [151]	81.2	80.8	77.8	77.8
QMF [73]	80.7	79.7	79.7	79.6
Proposed Model	**84.6** (↑ 3.4)	**84.5** (↑3.7)	**84.9** (↑ 4.7)	**91.1** (↑11.2)

Table 6.2: Effectiveness of content-level fusion approaches on CMU-MOSI and CMU-MOSEI. Best results are highlighted in boldface. Numbers in parentheses indicate relative percentage improvement over the next best model.

Table 6.3 shows the effectiveness of the existing decision-level fusion approaches on IEMOCAP task. Among the baseline decision-level fusion strategies, Stacking turned out to be the most effective for all emotion categories, except for the neutral category where Deep Fusion (see Talbe 6.3) approach achieved better performance. Overall, our proposed approach outperformed all the baseline fusion strategies. In particular, the proposed model achieved an increased accuracy of 92.9% as compared to 88.6% of Stacking for the happy class (i.e., 4.3% significant improvement), 89.6% as compared to 85.2% of Stacking for the sad class (i.e., 4.4% significant improvement), 90.3% as compared to 87.3% of Stacking for the angry class (i.e., 3.0% significant improvement), and 75.1% as compared to 70.9% of Deep Fusion for the neutral class (i.e., 4.2% significant improvement). Furthermore, the effectiveness of content-level fusion approaches on IEMOCAP is illustrated in Table 6.4. Again, our proposed method achieved a significant improvement of 2.8%, 3.4%, 2.5%, and 3.0% for the happy, sad, angry, and neutral classes, respectively, as compared to the next best models.

	Happy		Sad		Angry		Neutral	
Approach	Acc_2	$F1$	Acc_2	$F1$	Acc_2	$F1$	Acc_2	$F1$
Hard Voting	82.8	80.6	76.6	74.4	82.4	81.1	64.8	62.3
Weighted Voting	85.7	83.1	79.4	77.2	85.6	82.4	67.5	64.9
Soft Voting	83.9	81.4	77.7	75.1	83.5	81.7	65.8	63.2
SVM	86.5	83.8	80.8	78.9	85.9	82.6	68.4	65.2
Logistic Regression	87.2	84.3	81.5	80.1	86.7	83.5	69.1	65.9
GaussianNB	86.1	83.5	80.2	79.3	85.3	82.2	67.8	65.0
Stacking	88.6	86.4	85.2	84.1	87.3	84.5	70.6	67.3
Bagging	88.2	86.1	84.8	83.2	86.6	83.4	70.1	66.9
Boosting	87.7	85.5	84.1	82.6	85.9	82.5	69.6	66.4
Deep Fusion	86.8	84.2	81.3	80.0	86.2	83.1	70.9	67.8
Proposed Model	**92.9** (↑ 4.3)	**92.7** (↑ 6.3)	**89.6** (↑4.4)	**89.5** (↑ 5.4)	**90.3** (↑ 3.0)	**90.1** (↑ 5.6)	**75.1** (↑4.2)	**75.0** (↑ 7.2)

Table 6.3: Effectiveness of decision-level fusion approaches on IEMOCAP. Best results are highlighted in boldface. Numbers in parentheses indicate relative percentage improvement over the next best model.

	Happy		Sad		Angry		Neutral	
Approach	Acc_2	$F1$	Acc_2	$F1$	Acc_2	$F1$	Acc_2	$F1$
MulT [126]	90.1	88.2	86.2	85.6	87.3	87.0	72.1	70.3
RAVEN [143]	88.4	86.8	84.6	83.0	87.3	86.8	69.8	69.4
TFN [151]	87.5	85.6	83.7	82.8	87.8	87.0	70.3	69.6
Proposed Model	**92.9** (↑ 2.8)	**92.7** (↑ 4.5)	**89.6** (↑3.4)	**89.5** (↑ 3.9)	**90.3** (↑ 2.5)	**90.1** (↑ 3.1)	**75.1** (↑3.0)	**75.0** (↑ 4.7)

Table 6.4: Effectiveness of content-level fusion approaches on IEMOCAP. Best results are highlighted in boldface. Numbers in parentheses indicate relative percentage improvement over the next best model.

6.5.5 Ablation Tests

Table 6.5 shows the results of our ablation study. The first three rows list the performance of unimodal classifiers when no crossmodal interactions were modelled. The linguistic modality was the most predictive due to the use of word embedding trained on large corpora. For CMU-MOSEI, the

linguistic classifier outperformed all the content-level and voting-based fusion approaches, illustrating the robustness of uni-modal classifiers.

	CMU-MOSI		CMU-MOSEI	
Approach	Acc_2	$F1$	Acc_2	$F1$
Linguistic Only	77.1	72.3	81.5	87.8
Visual Only	54.7	48.4	71.1	83.0
Acoustic Only	56.1	60.0	71.2	83.1
Proposed Model	**84.6** (↑ 7.5)	**84.5** (↑12.2)	**84.9** (↑ 3.4)	**91.1** (↑3.3)

Table 6.5: Comparison with uni-modal sentiment analysis classifiers.

As a second set of ablation experiments, we tested the proposed model when only bimodal dynamics were present. We present the result in Table 6.6, which shows the linguistic and acoustic dynamics were the most informative. However, trimodal dynamics outperformed all possible bimodal combinations, yielding an improvement of accuracy by 5.0% for CMU-MOSI, and 2.2% for CMU-MOSEI.

	CMU-MOSI		CMU-MOSEI	
Approach	Acc_2	$F1$	Acc_2	$F1$
Model$_{\{Linguistic,Visual\}}$	78.2	74.3	82.1	88.4
Model$_{\{Linguistic,Acoustic\}}$	79.6	75.1	82.7	89.2
Model$_{\{Visual,Acoustic\}}$	55.1	55.2	70.8	82.7
Proposed Model	**84.6** (↑ 5.0)	**84.5** (↑9.4)	**84.9** (↑ 2.2)	**91.1** (↑1.9)

Table 6.6: Comparison of the model with its variants.

6.5.6 Effect of Incompatibility

We conducted a further analysis to investigate the effectiveness of incompatibility. We first identified all the cases that were correctly predicted by one out of the eleven decision-level fusion approaches. In total, there were 33 such cases on CMU-MOSI and 1547 ones on CMU-MOSEI test sets. The proposed model gave correct predictions for 31 cases out of 33 on CMU-MOSI and all the 1547 cases on CMU-MOSEI. Furthermore, we analyzed the cases that all uni-modal classifiers gave wrong sentiment judgments, but the proposed model successfully fused them giving correct predictions. There were 39 such utterances out of 686 on the CMU-MOSI and 633 utterances out of 4643 on the CMU-MOSEI subsets.

6.5.7 Case Study

We illustrate the visual-acoustic content of an incompatible case of the utterance *"I mean even if you don't have kinds"* in Figure 6.3. The linguistic state by itself is in an indefinite state, which

results in a superposition of sentiment judgments. Similarly, the visual-acoustic content is under uncertainty since the content is neutral. Indeed, all uni-modal classifiers predicted a negative sentiment judgment, inferring a probability less than 0.5, yet very close to the decision boundary of 0.5. This superposition of uni-modal beliefs, i.e., positive and negative sentiment at the same time until they are judged under a specific context, results in the occurrence of incompatibility. Under the high levels of uncertainty, incompatibility influences the uni-modal judgments and successfully predicts a positive multimodal sentiment judgment. This phenomenon is the core of the model and the reason it achieves such high performance.

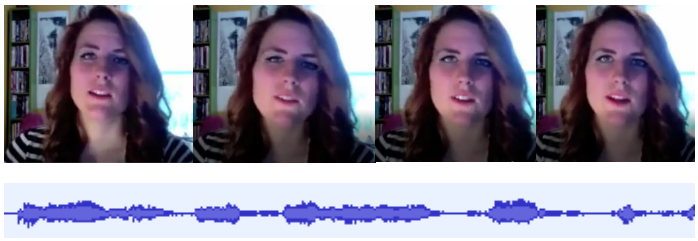


Figure 6.3: Visual-acoustic content of an incompatible case.

6.6 Conclusions

We have proposed an effective fusion strategy inspired by quantum cognition. Specifically, we formulated utterances as states and uni-modal decisions as mutually incompatible observables in a complex-valued sentimental Hilbert space. The incompatibility captures cognitive biases that are otherwise not possible with classical probability. The proposed model has been shown to be able to handle all combination patterns including the cases where all uni-modal classifiers gave wrong sentiment judgments. Therefore, the proposed approach achieved an improved performance over SOTA content-level and decision-level modality fusion approaches.

Chapter 7

Conclusions and Discussions

7.1 Conclusions

The main aim of the research reported in this thesis was to find out how the mathematical framework of Quantum Theory could be applied into workable computational models to accommodate rational and irrational cognitive biases, underlying the multimodal decision perspectives, and to yield a practical significance for analyzing human language. This has been broken down into four sub-research questions.

The first research question (**RQ 1**) is concerned about current SOTA fusion approaches and how specific components in those approaches facilitated solving the problem. We answered it in Chapter 2 by conducting a large-scale and comprehensive empirical comparison of SOTA modality fusion approaches for analyzing human language. The experiments showed that all neural models could not cope with skewed datasets. Additionally, the existing modality fusion approaches were data-hungry since the performance was lower for small-scale datasets as compared to large-scale datasets. Moreover, a more detailed analysis showed that SOTA models could not accommodate ambiguous cases. For instance, when an utterance itself was vague, and there was no specific context to bias its emotional state, neural models failed to make correct predictions. Finally, among the distinct modality fusion approaches, those that applied bi-modal attention mechanisms to model interactions across many modalities were the most effective. We expect that the findings did not only provide helpful insights for this dissertation to make steps forward but also they will provide guidance to other researchers to the development of more effective modality fusion models.

The next research question (**RQ 2**) focused on how we could effectively exploit the mathematical formalism of quantum mechanics outside physics to fuse inputs of multimodal features. We answered this question in Chapter 4. The work contributed to the field of quantum-inspired multimodal analysis in both theory and implementation. An end-to-end quantum-inspired framework for tackling a multimodal task was constructed, and a multimodal fusion was conducted down at

the word level. Furthermore, that model was the first to introduce complex values to implement the quantum process into the multimodal context on the implementation level. In addition to obtaining comparable performance to SOTA models, we also contributed an interpretation approach to facilitate understanding of multimodal interactions from both quantum and classical perspectives. In particular, the quantum view brought a unique superiority in understanding the contributions of single modalities and bi-modals (i.e., any of two modalities) to the predicted sentiment judgment, without the need for extra tuning.

Although our results were encouraging, the model in Chapter 4 was subject to multiple limitations. In particular, the interactions across distinct modalities were largely absent from the feature level. Moreover, the multimodal representation was a separable state rather an entangled representation with respect to the three uni-modal modalities. This means that the strategy did not fully exploit the expressive power of quantum probabilities to fuse inputs of multimodal features. That brought to the third research question (**RQ 3**): How could we encode cross-modal information in the form of non-classical correlations and how such correlations could benefit multimodal decision making? We answered that question in Chapter 5. The work contributed to the fields of quantum-inspired multimodal representation learning and affecting computing. In terms of quantum-inspired representation learning, we deployed an end-to-end quantum probabilistic neural model to address the non-separability of cross-modal interactions. To our best knowledge, no existing models in the current literature have taken into account. In addition to obtaining improved performance, we proposed a flexible neural architecture, which considered both context modelling, in the form of preceding utterances, and multimodal fusion, in the form of modality interactions, into a unified framework. Moreover, the model architecture supported multiparty conversations without requiring artificial expansion. We also contributed an interpretation approach by quantifying the degree of non-separability between modalities, unearthing useful and explainable knowledge about the way distinct modalities interacted with each other.

Although the work in Chapter 5 addressed various limitations in quantum-inspired multimodal representation learning field, it remained an open research question (**RQ 4**): how the quantum cognition could be precisely exploited to represent explicitly users' cognitive states, underlying their multimodal decision perspectives, and how we could learn such representations from multimodal data. We answered this question in Chapter 6 by proposing the very first quantum cognitively inspired theoretical approach, with practical implementations, that investigated and modelled the incompatibility of multimodal sentiment judgments for video sentiment analysis. We also introduced a novel methodology to construct the complex-valued Hilbert space representation, exploited to facilitate a decision fusion approach. In addition to obtaining improved performance, we contributed an explainable approach for decisions made, via the concept of incompatibility.

One limitation of the thesis concerns the visualization of abstract sentiment and emotion concepts, i.e., quantum measurements. Currently, the most effective way to visualize embeddings by

projecting high dimensional representations onto a 2-dimension space is the t-SNE (t-Distributed Stochastic Neighbouring Entities). However, the quantum-view of embeddings, i.e., complex-valued embeddings, and the learning process of the complex-valued neural network enable impractical such a visualization. In particular, we have trained the model by feeding the real and imaginary parts of the complex-valued layers as different input parts and simulated complex operations using real values. This means that each abstract concept conveys two representation parts. One might consider visualizing the real and imaginary part of the abstract sentiment concepts independently. However, we consider that such an approach (i.e., visualizing the real and imaginary parts separately) will result in clusters of abstract concepts, which do not reflect how our complex-valued neural model encodes sentiment and emotion information.

7.2 A Broader Discussion

Artificial neural networks have proven to be very efficient at detecting patterns in large sets of data, and they can do it in a scalable way. The experiments in Chapter 2 showed that increasing the size of neural networks and training them on larger sets of annotated data will, in most cases, improve their accuracy. This characteristic has created a sort of "bigger is better" mentality, pushing some AI researchers to seek improvements and breakthroughs by creating larger and larger AI models and datasets. Indeed, AI models now perform tasks like image classification, object detection and facial recognition with accuracy that often exceeds that of humans thanks to deep neural architectures and the existence of large datasets. However, current AI systems suffer from flaws that will not be fixed by making them bigger. For instance, an AI system trained to infer binary sentiment analysis of video utterances will not be able to do anything else, not even identify emotions that is a slightly different task. Moreover, neural network models are vulnerable to adversarial examples, perturbations in data that cause the AI systems to act in erratic ways. Finally, beyond the challenges of generalizability in AI models, most of them cannot overcome the "black-box problem", enabling them of limited utility in real-life applications.

In the past couple of years, there have been many discussions regarding the limits and challenges of deep learning, and there are various efforts into solving individual problems such as creating AI systems that are explainable, less data-hungry, and generalize well overcoming data biases. Some of the initiatives in the field involve the use of elements of symbolic artificial intelligence, the rule-based approach that dominated the field of AI before the rise of deep learning. However, this leads to a set of rules growing largely with the size of the data, which is impractical for real-world applications. On the contrary, there might need to go back to classical AI to deal with things like high-level cognition by injecting concepts from cognitive science into existing computational workable frameworks. For instance, Lin and He have proposed a probabilistic unsupervised model, which detects sentiment and topic of utterances, simultaneously [81]. Such approaches might be more similar to the way human

working memory processes information as compared to the SOTA supervised learning strategies. Currently, representation learning approaches are constructed in a way which is quite far from how humans understand and reason about cognitive states. Addressing such a research challenge could be beneficial for a great variety of representation learning tasks. First, the exploitation of a cognitive framework in the deployment of an AI model could emulate cognitive biases while taking decisions. Besides, it could draw attention to the various adversarial perturbations, which would lead to wrong predictions or decisions. Moreover, cognition could provide a spectrum of methodological tools to explain the decisions made. More importantly, it could give an insight into the causality established by the learning model as well as the reasoning of the model.

Causality is one element of higher-level cognition, which is a big debate these days. It is related to systematic generalization, which is the ability humans have to generalize the concepts they know, so they can be combined in new ways that are unlike anything else they have seen. That is a major concern of people working in deep learning and is very important for the next steps of progress of machine learning. In classical AI, researchers tried to obtain causality with logic and symbols. However, there are people like us who think that they shall inject functionalities into existing tools which have been built in the last few years. Such functionalities should be similar to the way humans perform reasoning, which is actually quite different from the way a purely logical system does it.

So far, deep learning has comprised learning from static datasets, which makes AI really good as tasks are related to correlations and associations. However, neural networks do not interpret cause and effect or why these associations and correlations exist. This, in turn, limits AI from being able to generalize their learning and transfer their skills to another related environment. In Chapter 5, we have leveraged the concept of non-separability, i.e., entanglement, not only to fuse multimodal information effectively but also to understand interactions across different modalities. Specifically, the bipartite Von Neumann entanglement entropy gave insights into the causality established by the learning model as well as the reasoning of the model. For instance, the decision of an ambiguous modality, e.g., visual, could be attributed due to entanglement with the informative modality, i.e., linguistic. Moreover, the effective degree of freedom contributing to the entanglement of bi-modals, via the calculation of Schmidt numbers, assisted in understanding the way that bi-modals interacted with each other and dealt with biases of training sub-datasets. Explicitly, the post-hoc analysis showed that the average Schmidt score of bi-modals was higher when a dataset was small, with many ambiguous cases. Contrary to machine learning approaches which discover correlations among data, the model in Chapter 5 exploits the notion of quantum entanglement, being a kind of non-classical and non-separable correlation, which provides a new mathematical formalism to unveil cause-effect relationships from correlations [160]. Crucially, quantum probabilities, as a more general and flexible probability theory, do not require the need to know *apriori*, whilst machine learning approaches require a wide frame of prior knowledge to prove that observed effects are casual.

Another crucial challenge in AI is how we can create functions similar to human reasoning.

Attention mechanisms were a potential strategy to address the above research question. In particular, attention mechanisms allow us to learn how to focus our computation on particular elements. For humans, attention is also an important part of conscious processing. For example, when someone is conscious of something, she/he focuses on a certain thought, and then she/he might move on to another thought. This is very different from standard neural networks, which are instead parallel processing on a large scale. Having said that, we have been experiencing big breakthroughs on computer vision, translation, and memory thanks to attention mechanisms.

Although the ability of neural networks to parallel-process on a large scale has given us significant breakthroughs, research is now shifting to developing novel deep architectures and training frameworks for addressing tasks like reasoning, planning, capturing causality, and obtaining systematic generalization. In particular, there is a need for investigating computational cognitive models that involve the human in the loop and thereby, become human-centric. A good starting point towards this research direction is the work of Piwek et al. [108], who proposed a cognitive model to investigate the impact of distal and proximal demonstratives in language comprehension. However, the analysis revealed that the relation between proximals/distals and importance was refuted. A post-hoc analysis showed that importance might be linked with the use of a pointing act rather than the choice of demonstrative.

On the other side, clearly, the application of QT beyond its standard domain could be controversial, but it can identify some intriguing and fresh directions for computational modelling. Crucially, QT provides a robust mathematical framework beyond its physical theory, in which various theories can be developed. We argue that the quantum formalism of quantum mechanics and quantum cognition are a different style of brain-inspired computations and a robust tool to get started towards the new research trends. To this end, in Chapter 6, we have introduced a quantum cognitively motivated framework to model the fusion of decisions from multimodal decision perspectives via the concept of incompatibility. The modelling of uni-modal decisions as mutually incompatible is able to capture not only rational and optimal decisions, but also sub-optimal, irrational, and paradoxical decisions. Practically, this means that the concept of incompatibility, accommodated only by quantum probabilities, allows effective handling of all combination patterns, including those extreme cases that are wrongly predicted by all uni-modal classifiers.

Another step that will help AI systems to behave more consistently is how they decompose data and find the important bits. Attention mechanisms, which enable neural networks to focus on relevant bits of information, and transfer learning, mapping the parameters of one neural network to another, represent a great progress towards this direction. Nevertheless, a better compositionality can lead to deep learning systems that can extract and manipulate high-level features in their problem domains and dynamically adapt them to new environments without the need for extra tuning and lots of data. In line with this, in Chapter 4, we have proposed a quantum-inspired multimodal fusion model for video sentiment analysis. The quantum view of the multimodal representation learning

has brought a unique superiority in understanding the contributions of uni-modal and bi-modal dynamics to the predicted final sentiment judgment, via the concept of the reduced matrix, without requiring to retrain the neural model.

It is customary to think that by focusing solely on performance, the models will be increasingly opaque. This is true in the sense that there is a trade-off between performance and model transparency. However, the experiments in Chapters 4, 5, and 6 showed the benefits of exploiting the mathematical formalism of quantum mechanics to fuse different modalities. In particular, the experiments showed the effectiveness of quantum probabilistic models achieving comparable or superior performance as compared to various SOTA fusion strategies. Additionally, they do unearth not only useful and explainable knowledge about the way distinct modalities interact with each other but also demonstrate high-level model transparency due to their theoretical root on the well-established quantum physics meaning.

7.3 Future Directions

In this dissertation, our motivation to use the mathematical formalism of quantum mechanics was to overcome deficiencies of explainability and cognitive biases in the existing fusion approaches. However, there is the need of other critical modelling aspects that should be taken into account when deploying AI-based systems in practice. That includes not only algorithmic proposals but also new procedures devoted to ensuring responsibility in the application and usage of AI models, including tools for accountability and data governance, methods to assess and explain the impact of decisions made by AI models, or techniques to detect, counteract or mitigate the effect of bias on the model's output. It is only by carefully accounting for all these aspects when humans, through all processes and systems endowed with AI-based functionalities (e.g. Robotics, Machine Learning, Optimization and Reasoning), will fully trust and welcome the arrival of this technology. Under this scope, quantum cognition could constitute a robust mathematical tool to developing methods for describing risks in AI applications in the future.

One major concern among the risks in AI applications is the privacy and model confidentiality in data fusion contexts. Computational representation learning models shall have complex representations of their learnt patterns. Not being able to understand what has been captured by the model and stored in its internal representation may entail a privacy breach. Contrarily, the ability to explain the inner relations of a trained model by non-authorized third parties may also compromise the differential privacy of the data origin. QT and quantum cognition could yield a spectrum of methodological approaches and mathematical tools establishing a solid research ground towards this direction. Indeed, QT has been harnessed for use in a variety of real-world applications, such as cryptography. Currently, quantum cryptography is considered the most secure due to the unbreakable quantum key distribution. In this dissertation, we have leveraged the notion of entanglement and

incompatibility to unveil cause-effect relationships among distinct modalities. However, the same quantum properties and the bizarre behaviour of QT are worth being investigated for addressing safety and security issues over fusion processes.

Another concern is that AI models should be able to generalize efficiently and to a large scale, and handle the uncertainties of the world. Currently, AI models have to be trained anew when a slight change is brought to their environment. AI models need to generalize to different distributions in data and do continual learning. For instance, if someone has learnt driving in Greece, she/he does not need to learn driving all over again when she/he moves to the UK. She/he just have to adapt herself/himself to the new environment. Similarly, AI models should be able to handle environment changes around them. This is a long term goal and there is no solution yet. To this end, the properties of open quantum systems could be a possible direction to explore. In particular, an open quantum system is a quantum-mechanical system that interacts with an external quantum system, which is known as the *environment* or an *ensamble*. In general, such interactions significantly change the dynamics of the system and result in quantum dissipation, such that the information contained in the system is lost to its environment. In Section 5.2, we introduced a methodology to model interactions between current and preceding utterances via quantum evolution. We viewed the current utterance like a particle which interacts with an external environment of particles, corresponding to preceding utterances. In the future, open quantum systems in conjunction with quantum properties, like superposition, could be leveraged to manipulate high-level features in their problem domains and dynamically adapt them to new environments, without the need for extra tuning. However, more sophisticated ways of interactions than those in Section 5.2 are required. To make further progress, it is these kinds of problems that we must turn to next.

Appendix A

Fine-tuning Final Settings of Baselines

Tables list the final settings for each neural model, after the fine-tuned process through a fifty-times random grid search on the hyper-parameters.

Table A.1: Hyperparameters of EF-LSTM on CMU-MOSI, CMU-MOSEI, and IEMOCAP.

	CMU-MOSI	CMU-MOSEI	IEMOCAP
Batch Size	64	64	16
Initial Learning Rate	0.002	0.002	0.001
LSTM Output	96	128	128
Multimodal Embedding Dimension	64	128	16
Multimodal Embedding Dropout	0.1	0.2	0.1
Gradient Glip	0.4	0.8	0.3

Table A.2: Hyperparameters of LF-LSTM on CMU-MOSI, CMU-MOSEI, and IEMOCAP.

	CMU-MOSI	CMU-MOSEI	IEMOCAP
Batch Size	64	16	16
Initial Learning Rate	0.005	0.001	0.001
LSTM Outputs	128,16,80	128,64,16	128,64,16
Multimodal Embedding Dimension	32	48	32
Multimodal Embedding Dropout	0.2	0.4	0.2
Gradient Glip	0.4	0.3	0.7

Table A.3: Hyperparameters of TFN on CMU-MOSI, CMU-MOSEI, and IEMOCAP.

	CMU-MOSI	CMU-MOSEI	IEMOCAP
Batch Size	96	128	128
Initial Learning Rate	0.001	0.002	0.001
Subnetwork Outputs	128,80,80	128,16,32	128,80,60
Subnetwork Dropout Probabilities	0.1,0.1,0.1	0.2,0.2,0.2	0.5,0.5,0.5
Sentiment Subnetwork Outpout	16	96	128
Sentiment Subnetwork Probability	0.4	0.3	0.4
Gradient Glip	0.1	0.1	0.5

Table A.4: Hyperparameters of LMF on CMU-MOSI, CMU-MOSEI, and IEMOCAP.

	CMU-MOSI	CMU-MOSEI	IEMOCAP
Batch Size	96	128	32
Initial Learning Rate	0.001	0.002	0.001
Rank	4	4	16
Subnetwork Outputs	128,32,80	128,64,32	128,64,32
Subnetwork Dropout Probabilities	0.5,0.5,0.5	0.1,0.1,0.1	0.3,0.3,0.3
Crossmodal Representation	0.2	0.2	0.4
Gradient Glip	0.2	0.2	0.4

Table A.5: Hyperparameters of MARN on CMU-MOSI, CMU-MOSEI, and IEMOCAP.

	CMU-MOSI	CMU-MOSEI	IEMOCAP
Batch Size	128	16	64
Initial Learning Rate	0.001	0.002	0.001
LSTM Outputs	128,64,80	128,80,80	128,80,32
Attention Blocks	2	2	5
Attention Cell	16	64	32
Compressed dimension	64,32,8	64,40,40	64,16,8
Output cell dimension	16	16	96
Gradient Glip	0.1	0.2	0.7

Table A.6: Hyperparameters of MFN on CMU-MOSI, CMU-MOSEI, and IEMOCAP.

	CMU-MOSI	CMU-MOSEI	IEMOCAP
Batch Size	128	128	32
Initial Learning Rate	0.001	0.002	0.005
LSTM Outputs	128,80,16	128,80,16	128,64,16
γ_1, γ_2 cell dimensions	128,128	128,128	64,32
Attention cell dimensions	64,32	64,32	256,32
Memory dimension	256	256	256
Output cell dimension	64	64	128
Gradient Glip	0.2	0.2	0.7

Table A.7: Hyperparameters of MulT on CMU-MOSI, CMU-MOSEI, and IEMOCAP.

	CMU-MOSI	CMU-MOSEI	IEMOCAP
Batch Size	16	128	32
Initial Learning Rate	0.001	0.002	0.005
Transformers Hidden Unit Size	40	40	40
Crossmodal Blocks	4	4	4
Crossmodal Attention Heads	8	10	10
Temporal Convolution Kernel Size	3/3/3	3/3/3	3/3/5
Textual Embedding Dropout	0.3	0.2	0.3
Crossmodal Attention Block Dropout	0.1	0.2	0.25
Output Dropout	0.1	0.1	0.1
Gradient Glip	0.2	0.2	0.7

Table A.8: Hyperparameters of MMUU-BA on CMU-MOSI, CMU-MOSEI, and IEMOCAP.

	CMU-MOSI	CMU-MOSEI	IEMOCAP
Batch Size	64	64	64
Initial Learning Rate	0.005	0.002	0.001
RNN dropouts	0.15,0.15,0.15	0.1,0.1,0.1	0.7,0.7,0.7
GRU dropouts	0.1,0.1,0.1	0.3,0.3,0.3	0.15,0.15,0.15
FC dropouts	0.15,0.15,0.15	0.8,0.8,0.8	0.15,0.15,0.15
Output cell dimensions	32	32	64
Output dropout	0.15	0.3	0.1
Gradient Glip	0.3	0.9	0.5

Table A.9: Hyperparameters of RMFN on CMU-MOSI, CMU-MOSEI, and IEMOCAP.

	CMU-MOSI	CMU-MOSEI	IEMOCAP
Batch Size	64	128	16
Initial Learning Rate	0.005	0.002	0.002
Shift Weight	0.8	0.7	0.1
LSTM layers	3	1	1
Cell Output	50	40	30
Gradient Glip	0.7	1	0.1

Table A.10: Hyperparameters of RMFN on CMU-MOSI, CMU-MOSEI, and IEMOCAP.

	CMU-MOSI	CMU-MOSEI	IEMOCAP
Batch Size	64	128	16
Initial Learning Rate	0.005	0.002	0.002
Shift Weight	0.8	0.7	0.1
LSTM layers	3	1	1
Cell Output	50	40	30
Gradient Glip	0.7	1	0.1

Appendix B

Expectation Values of Observables

The expectation of a random variable X that takes the values $\{+,-\}$ according to the probability distribution $P_{X(+)}, P_{X(-)}$ is defined as

$$\langle X \rangle = (+)P_{X(+)} + (-)P_{X(-)}.$$

For two random variables X, Ψ, that take the values $\{+,-\}$ according to the probability distribution $P_{X(+)}, P_{X(-)}$ and $P_{\Psi(+)}, P_{\Psi(-)}$ respectively, the expectation value is defined as the product resulting in

$$\begin{aligned}\langle X, \Psi \rangle &= ((+)P_{X(+)} + (-)P_{X(-)}) * ((+)P_{\Psi(+)} + (-)P_{\Psi(-)}) \\ &= (+)(+)(P_{X(+)}P_{\Psi(+)}) + (+)(-)(P_{X(+)}P_{\Psi(-)}) \\ &+ (-)(+)(P_{X(-)}P_{\Psi(+)}) + (-)(-)(P_{X(-)}P_{\Psi(-)}).\end{aligned}$$

Appendix C

Correlations of Observables

Suppose M be an observable corresponding to one modality:

$$M = (+1)[M, +] + (-1)[M, -]$$

where $[M, .] = |M, .\rangle \langle M, .|$. If M is treated as random variable, it is possible to express its expectation as

$$\mathrm{E}(M) = \mathrm{tr}(\rho M) = (+1)P_m(+1) + (-1)P_m(-1)$$

where $\mathrm{tr}(\rho[M, +]) = P_m(+1)$ and $\mathrm{tr}(\rho[M, -]) = P_m(-1)$. Note that the variables are still real because the observed values, which are the eigenvalues of the observables, are real although the operators are defined over the complex field.

The variance of M can be expressed as

$$\begin{aligned} \mathrm{V}(M) &= (+1 - \mathrm{E}(M))^2 P_m(+1) \\ &+ (-1 - \mathrm{E}(M))^2 P_m(-1) \\ &= 1 - \mathrm{E}(M)^2 \end{aligned}$$

We can assume that observable expectations are zero without loss of generality. If they were not zero, it is possible to normalize the observables and define

$$\bar{M} = (+1 - \mathrm{E}(M))[M, +] + (-1 - \mathrm{E}(M))[M, -1] \ .$$

such that

$$\mathrm{E}(\bar{M}) = \mathrm{E}(M - \mathrm{E}(M)) = \mathrm{E}(M) - \mathrm{E}(\mathrm{E}(M)) = 0$$

As a consequence, suppose $\bar{M} \equiv M$ without loss of generality. Therefore, $\mathrm{V}(M) = 1$.

Suppose A, B be two observables corresponding to two modalities. Both variables have zero expectation and unit variance. The covariance between A, B can be expressed as the following complex number:

$$\mathrm{Cov}(A,B) = \sum_{i\in\{-1,+1\}} \sum_{j\in\{-1,+1\}} (i - \mathrm{E}(A))(j - \mathrm{E}(B)) P_{a,b}(i,j)$$

where

$$P_{a,b}(i,j) = \frac{\mathrm{tr}([A,i]\rho[A,i][B,j])}{\mathrm{tr}([A,i]\rho[A,i])}$$

is the probability that A is measured after the measurement of B according to Lüders' rule [86]. The covariance between A, B can be expressed in the following expanded form since both expectations are zero:

$$\begin{aligned}
\mathrm{Cov}(A,B) \quad = \quad & \frac{\mathrm{tr}([A,+1]\rho[A,+1][B,+1])}{\mathrm{tr}([A,+1]\rho[A,+1])} + \\
& \frac{\mathrm{tr}([A,+1]\rho[A,+1][B,-1])}{\mathrm{tr}([A,+1]\rho[A,+1])} - \\
& \frac{\mathrm{tr}([A,-1]\rho[A,-1][B,+1])}{\mathrm{tr}([A,-1]\rho[A,-1])} - \\
& \frac{\mathrm{tr}([A,-1]\rho[A,-1][B,-1])}{\mathrm{tr}([A,-1]\rho[A,-1])}
\end{aligned}$$

The correlation between two any variables A, B is defined as

$$\mathrm{Corr}(A,B) = \frac{\mathrm{Cov}(A,B)}{\sqrt{\mathrm{V}(A)}\sqrt{\mathrm{V}(B)}}$$

that is

$$\mathrm{Corr}(A,B) = \mathrm{Cov}(A,B)$$

since $V(M) = 1$ for $M \in \{A, B\}$.

Then, consider the correlation between A, B and of the pure state $\rho = |g\rangle\langle g|$. We have that

$$\begin{aligned}
P_{a,b}(i,j) \quad &= \quad \frac{\mathrm{tr}([A,i]\rho[A,i][B,j])}{\mathrm{tr}([A,i]\rho[A,i])} \\
&= \quad \frac{\mathrm{tr}(|A,i\rangle\langle A,i|\,\rho\,|A,i\rangle\langle A,i|\,|B,j\rangle\langle B,j|)}{\mathrm{tr}(|A,i\rangle\langle A,i|\,\rho\,|A,i\rangle\langle A,i|)} \\
&= \quad \frac{\langle B,j|A,i\rangle\langle A,i|g\rangle\langle g|A,i\rangle\langle A,i|B,j\rangle}{\langle A,i|A,i\rangle\langle A,i|g\rangle\langle g|A,i\rangle} \\
&= \quad \frac{|\langle B,j|A,i\rangle|^2 |\langle A,i|g\rangle|^2}{|\langle A,i|g\rangle|^2} \\
&= \quad |\langle B,j|A,i\rangle|^2
\end{aligned}$$

It follows that

$$\begin{aligned} \mathrm{Corr}(A,B) &= |\langle B,+1|A,+1\rangle|^2 + \\ &\quad |\langle B,-1|A,-1\rangle|^2 - \\ &\quad |\langle B,+1|A,-1\rangle|^2 - \\ &\quad |\langle B,-1|A,+1\rangle|^2 \end{aligned}$$

C.1 Hyperparameters of Uni-modal Classifiers

For uni-modal classifiers, the Bi-directional GRUs had 300 neurons, each followed by a dense layer consisting of 100 neurons. A grid search is conducted over a hyperparameter pool. We report the final settings in Table C.1. Each uni-modal classifiers was trained for 50 epochs with Adam as the optimizer on the L_1 loss function.

	CMU-MOSI			CMU-MOSEI		
Hyperparameter	L	V	A	L	V	A
Learning Rate	0.002	0.002	0.002	0.001	0.002	0.001
Batch Size	128	128	128	128	32	32
Gradient Clipping	0.8	0.5	0.8	0.3	1	0.5
Output	64	32	128	128	64	64
Output Dropout	0.15	0.2	0.3	0.3	0.5	0.4
GRU Dropout	0.2	0.5	0.1	0.3	0.2	0.5

Table C.1: Final settings for training uni-modal classifiers on CMU-MOSI and CMU-MOSEI

Bibliography

[1] Scott Aaronson. *Quantum Computing Since Democritus*. Cambridge University Press, 2013.

[2] Diederik Aerts. Quantum structure in cognition. *Journal of Mathematical Psychology*, 53(5):314–348, 2009.

[3] Diederik Aerts and Marek Czachor. Quantum aspects of semantic analysis and symbolic artificial intelligence. *Journal of Physics A: Mathematical and General*, 37(12):L123, 2004.

[4] Diederik Aerts, Suzette Geriente, Catarina Moreira, and Sandro Sozzo. Testing ambiguity and machina preferences within a quantum-theoretic framework for decision-making. *Journal of Mathematical Economics*, 78:176–185, 2018.

[5] Diederik Aerts and Sandro Sozzo. Quantum structure in cognition: Why and how concepts are entangled. In *International Symposium on Quantum Interaction*, pages 116–127. Springer, 2011.

[6] Diederik Aerts and Sandro Sozzo. Quantum entanglement in concept combinations. *International Journal of Theoretical Physics*, 53(10):3587–3603, 2014.

[7] Hameed M Al Janaby and Ammar A Abed. Syntactic ambiguity in newspaper headlines. *Journal of the Faculty of Collective Knowledge*, 16:1–29, 2011.

[8] Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. In *International Conference on Machine Learning*, pages 1120–1128, 2016.

[9] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[10] Alain Aspect, Philippe Grangier, and Gérard Roger. Experimental realization of einstein-podolsky-rosen-bohm gedankenexperiment: a new violation of bell's inequalities. *Physical review letters*, 49(2):91, 1982.

[11] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379, 2010.

[12] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

[13] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[14] Leonardo Banchi, Edward Grant, Andrea Rocchetto, and Simone Severini. Modelling non-markovian quantum processes with recurrent neural networks. *New Journal of Physics*, 20(12):123030, 2018.

[15] Elham J Barezi and Pascale Fung. Modality-based factorization for multimodal fusion. *arXiv preprint arXiv:1811.12624*, 2018.

[16] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[17] Peter Bruza and Vivien Chang. Perceptions of document relevance. *Frontiers in psychology*, 5:612, 2014.

[18] Peter Bruza, Kirsty Kitto, and Doug McEvoy. Entangling words and meaning. In *Quantum Interaction: Proceedings of the Second Quantum Interaction Symposium (QI-2008):*, pages 118–124. College Publications, 2008.

[19] Peter Bruza, Kirsty Kitto, Douglas Nelson, and Cathy McEvoy. Is there something quantum-like about the human mental lexicon? *Journal of Mathematical Psychology*, 53(5):362–377, 2009.

[20] Peter D Bruza and Richard J Cole. Quantum logic of semantic space: An exploratory investigation of context effects in practical reasoning. *arXiv preprint quant-ph/0612178*, 2006.

[21] Peter D Bruza, Kirsty Kitto, Brentyn Ramm, Laurianne Sitbon, Dawei Song, and Simon Blomberg. Quantum-like non-separability of concept combinations, emergent associates and abduction. *Logic Journal of IGPL*, 20(2):445–457, 2012.

[22] Peter D Bruza, Kirsty Kitto, Brentyn J Ramm, and Laurianne Sitbon. A probabilistic framework for analysing the compositionality of conceptual combinations. *Journal of Mathematical Psychology*, 67:26–38, 2015.

[23] Curt Burgess, Kay Livesay, and Kevin Lund. Explorations in context space: Words, sentences, discourse. *Discourse Processes*, 25(2-3):211–257, 1998.

[24] Paul Busch. Unsharp reality and joint measurements for spin observables. *Physical Review D*, 33(8):2253, 1986.

[25] Jerome R Busemeyer and Peter D Bruza. *Quantum models of cognition and decision*. Cambridge University Press, 2012.

[26] Jerome R Busemeyer and Zheng Wang. Hilbert space multidimensional theory. *Psychological review*, 125(4):572, 2018.

[27] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335, 2008.

[28] Erik Cambria. Affective computing and sentiment analysis. *IEEE intelligent systems*, 31(2):102–107, 2016.

[29] Erik Cambria, Newton Howard, Jane Hsu, and Amir Hussain. Sentic blending: Scalable multimodal fusion for the continuous interpretation of semantics and sentics. In *2013 IEEE symposium on computational intelligence for human-like intelligence (CIHLI)*, pages 108–117. IEEE, 2013.

[30] Iti Chaturvedi, Ranjan Satapathy, Sandro Cavallari, and Erik Cambria. Fuzzy commonsense reasoning for multimodal sentiment analysis. *Pattern Recognition Letters*, 125:264–270, 2019.

[31] Hao Chen. Schmidt numbers of low-rank bipartite mixed states. *Physical Review A*, 67(6):062301, 2003.

[32] Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 163–171, 2017.

[33] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[34] Boris S Cirel'son. Quantum generalizations of bell's inequality. *Letters in Mathematical Physics*, 4(2):93–100, 1980.

[35] John F Clauser, Michael A Horne, Abner Shimony, and Richard A Holt. Proposed experiment to test local hidden-variable theories. *Physical review letters*, 23(15):880, 1969.

[36] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarep—a collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 960–964. IEEE, 2014.

[37] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[38] Sri Harsha Dumpala, Rupayan Chakraborty, and Sunil Kumar Kopparapu. Audio-visual fusion for sentiment classification using cross-modal autoencoder. In *32nd Conference on Neural Information Processing Systems (NIPS 2018)*, pages 1–4, 2019.

[39] Ehtibar N Dzhafarov, Janne V Kujala, and Victor H Cervantes. Contextuality-by-default: a brief overview of ideas, concepts, and terminology. In *International Symposium on Quantum Interaction*, pages 12–23. Springer, 2015.

[40] Albert Einstein, Boris Podolsky, and Nathan Rosen. Can quantum-mechanical description of physical reality be considered complete? *Physical review*, 47(10):777, 1935.

[41] Lauren Fell, Shahram Dehdashti, Peter Bruza, and Catarina Moreira. An experimental protocol to derive and validate a quantum model of decision-making. *arXiv preprint arXiv:1908.07935*, 2019.

[42] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[43] Ingo Frommholz, Birger Larsen, Benjamin Piwowarski, Mounia Lalmas, Peter Ingwersen, and Keith Van Rijsbergen. Supporting polyrepresentation in a quantum-inspired geometrical retrieval framework. In *Proceedings of the third symposium on Information interaction in context*, pages 115–124, 2010.

[44] Liane Gabora and Diederik Aerts. Contextualizing concepts using a mathematical generalization of the quantum formalism. *Journal of Experimental & Theoretical Artificial Intelligence*, 14(4):327–358, 2002.

[45] Lorenzo Gatti, Marco Guerini, and Marco Turchi. Sentiwords: Deriving a high precision and high coverage lexicon for sentiment analysis. *IEEE Transactions on Affective Computing*, 7(4):409–421, 2015.

[46] Efthymios Georgiou, Charilaos Papaioannou, and Alexandros Potamianos. Deep hierarchical fusion with application in sentiment analysis. In *INTERSPEECH*, pages 1646–1650, 2019.

[47] Deepanway Ghosal, Md Shad Akhtar, Dushyant Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. Contextual inter-modal attention for multi-modal sentiment analysis. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3454–3466, 2018.

[48] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*, 2019.

[49] Dimitris Gkoumas, Qiuchi Li, Christina Lioma, Yijun Yu, and Dawei Song. What makes the difference? an empirical comparison of fusion strategies for multimodal language analysis. *Information Fusion*, 2020.

[50] Dimitris Gkoumas, Qiuchi Li, Massimo Melucci, and Song Dawei. A quantum cognitively motivated decision fusionframework for video sentiment analysis. In *AAAI (Under Review)*, 2021.

[51] Dimitris Gkoumas, Qiuchi Li, Massimo Melucci, Nie Jian-Yun, and Song Dawei. An entanglement-driven neural network model for contextual and non-separable modality fusion in conversational emotion recognition. In *Information Fusion (Under Review)*, 2021.

[52] Dimitris Gkoumas, Dawei Song, Qiuchi Li, and Massimo Melucci. An entanglement-driven fusion neural network for video sentiment analysis. In *The Web Conference 2021 (Under Review)*, 2021.

[53] Dimitris Gkoumas, Sagar Uprety, and Dawei Song. Investigating non-classical correlations between decision fused multi-modal documents. In *International Symposium on Quantum Interaction*, pages 163–176. Springer, 2018.

[54] Michael Glodek, Stephan Tschechne, Georg Layher, Martin Schels, Tobias Brosch, Stefan Scherer, Markus Kächele, Miriam Schmidt, Heiko Neumann, Günther Palm, et al. Multiple classifier systems for the classification of audio-visual emotional states. In *International Conference on Affective Computing and Intelligent Interaction*, pages 359–368. Springer, 2011.

[55] Stephen Grossberg. Nonlinear neural networks: Principles, mechanisms, and architectures. *Neural networks*, 1(1):17–61, 1988.

[56] Michael Grubinger, Paul Clough, Allan Hanbury, and Henning Müller. Overview of the imageclefphoto 2007 photographic retrieval task. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 433–444. Springer, 2007.

[57] Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. Multimodal affective analysis using hierarchical attention strategy with word-level alignment. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, page 2225. NIH Public Access, 2018.

[58] P.R. Halmos. *Finite-Dimensional Vector Spaces*. Undergraduate Texts in Mathematics. Springer, New York, USA, 1987.

[59] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. Icon: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2594–2604, 2018.

[60] Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2018, page 2122. NIH Public Access, 2018.

[61] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. *arXiv preprint arXiv:2005.03545*, 2020.

[62] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[63] Andreas Holzinger, Peter Kieseberg, Edgar Weippl, and A Min Tjoa. Current advances, trends and challenges of machine learning and knowledge extraction: from machine learning to explainable ai. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 1–8. Springer, 2018.

[64] R.I.G. Hughes. *The Structure and Interpretation of Quantum Mechanics*. Harvard University Press, Cambridge, MA, USA, 1989.

[65] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012.

[66] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific, 2013.

[67] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[68] Andreui Nikolaevich Kolmogorov and Albert T Bharucha-Reid. *Foundations of the theory of probability: Second English Edition*. Courier Dover Publications, 2018.

[69] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[70] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[71] Franck Laloë. Do we really understand quantum mechanics? strange correlations, paradoxes, and theorems. *American Journal of Physics*, 69(6):655–701, 2001.

[72] Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. Combining language and vision with a multimodal skip-gram model. *arXiv preprint arXiv:1501.02598*, 2015.

[73] Ji-An Li, Daoyi Dong, Zhengde Wei, Ying Liu, Yu Pan, Franco Nori, and Xiaochu Zhang. Quantum reinforcement learning during human decision-making. *Nature Human Behaviour*, 4(3):294–307, 2020.

[74] Jingfei Li, Peng Zhang, Dawei Song, and Yuexian Hou. An adaptive contextual quantum language model. *Physica A: Statistical Mechanics and its Applications*, 456:51–67, 2016.

[75] Qiuchi Li, Dimitris Gkoumas, Christina Lioma, and Massimo Melucci. Quantum-inspired multimodal fusion for video sentiment analysis. *Information Fusion*, 65:58–71, 2020.

[76] Qiuchi Li, Dimitris Gkoumas, Alessandro Sordoni, Jian-Yun Nie, and Massimo Melucci. Quantum-inspired neural network for conversational emotion recognition. In *AAAI 2021 (Under Review)*, 2021.

[77] Qiuchi Li, Sagar Uprety, Benyou Wang, and Dawei Song. Quantum-inspired complex word embedding. *arXiv preprint arXiv:1805.11351*, 2018.

[78] Qiuchi Li, Benyou Wang, and Massimo Melucci. Cnm: An interpretable complex-valued network for matching. *arXiv preprint arXiv:1904.05298*, 2019.

[79] Paul Pu Liang, Zhun Liu, Yao-Hung Hubert Tsai, Qibin Zhao, Ruslan Salakhutdinov, and Louis-Philippe Morency. Learning representations from imperfect time series data via tensor rank regularization. *arXiv preprint arXiv:1907.01011*, 2019.

[80] Paul Pu Liang, Ziyin Liu, Amir Zadeh, and Louis-Philippe Morency. Multimodal language analysis with recurrent multistage fusion. *arXiv preprint arXiv:1808.03920*, 2018.

[81] Chenghua Lin and Yulan He. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 375–384, 2009.

[82] Zachary C Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.

[83] Meng Liu, Liqiang Nie, Meng Wang, and Baoquan Chen. Towards micro-video understanding by joint sequential-sparse modeling. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 970–978, 2017.

[84] Zhen-Tao Liu, Qiao Xie, Min Wu, Wei-Hua Cao, Dan-Yun Li, and Si-Han Li. Electroencephalogram emotion recognition based on empirical mode decomposition and optimal feature selection. *IEEE Transactions on Cognitive and Developmental Systems*, 11(4):517–526, 2018.

[85] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*, 2018.

[86] Gerhart Lüders. Concerning the state-change due to the measurement process. *Annals of Physics*, 15(9):663–670, 2006.

[87] Kevin Lund and Curt Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research methods, instruments, & computers*, 28(2):203–208, 1996.

[88] Mark J Machina. Risk, ambiguity, and the rank-dependence axioms. *American Economic Review*, 99(1):385–92, 2009.

[89] Sijie Mai, Haifeng Hu, and Songlong Xing. Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 481–492, 2019.

[90] Sijie Mai, Haifeng Hu, and Songlong Xing. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 164–172, 2020.

[91] Sijie Mai, Songlong Xing, and Haifeng Hu. Locally confined modality fusion network with a global perspective for multimodal human affective computing. *IEEE Transactions on Multimedia*, 22(1):122–137, 2019.

[92] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6818–6825, 2019.

[93] Massimo Melucci. Towards modeling implicit feedback with quantum entanglement. In *QI*, Oxford, UK, 2008. College Publications.

[94] Massimo Melucci. *Introduction to information retrieval and quantum mechanics*. Springer, 2015.

[95] T Mikolov, M Karafiát, and L Burget. J. ˇcernocky, and s. khudanpur. recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*, 2010.

[96] Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 1045–1048. ISCA, 2010.

[97] Oskar Morgenstern. The theory of games. *Scientific American*, 180(5):22–25, 1949.

[98] Emilie Morvant, Amaury Habrard, and Stéphane Ayache. Majority vote of diverse classifiers for late fusion. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 153–162. Springer, 2014.

[99] Michael A Nielsen and Isaac Chuang. Quantum computation and quantum information, 2002.

[100] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, USA, 10th edition, 2011.

[101] Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 284–288, 2016.

[102] Desmond C Ong, Jamil Zaki, and Noah D Goodman. Computational models of emotion inference in theory of mind: A review and roadmap. *Topics in cognitive science*, 11(2):338–357, 2019.

[103] Anirban Pathak. *Elements of quantum computation and quantum communication*. Taylor & Francis, 2013.

[104] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[105] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[106] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6892–6899, 2019.

[107] Phuong Pham and Jingtao Wang. Predicting learners' emotions in mobile MOOC learning via a multimodal intelligent tutor. In Roger Nkambou, Roger Azevedo, and Julita Vassileva, editors, *Intelligent Tutoring Systems - 14th International Conference, ITS 2018, Montreal, QC, Canada, June 11-15, 2018, Proceedings*, volume 10858 of *Lecture Notes in Computer Science*, pages 150–159. Springer, 2018.

[108] Paul Piwek, Robbert-Jan Beun, and Anita Cremers. 'proximal'and 'distal'in language and cognition: Evidence from deictic demonstratives in dutch. *Journal of Pragmatics*, 40(4):694–718, 2008.

[109] Benjamin Piwowarski, Ingo Frommholz, Mounia Lalmas, and Keith Van Rijsbergen. What can quantum theory bring to information retrieval. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 59–68, 2010.

[110] Norman Poh and Samy Bengio. How do correlation and variance of base-experts affect fusion in biometric authentication tasks? *IEEE Transactions on Signal Processing*, 53(11):4384–4396, 2005.

[111] Moacir P Ponti Jr. Combining classifiers: from the creation of ensembles to the decision fusion. In *2011 24th SIBGRAPI Conference on Graphics, Patterns, and Images Tutorials*, pages 1–10. IEEE, 2011.

[112] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017.

[113] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 873–883, 2017.

[114] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 439–448. IEEE, 2016.

[115] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018.

[116] Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Erik Cambria, Alexander Gelbukh, and Amir Hussain. Multimodal sentiment analysis: Addressing key issues and setting up the baselines. *IEEE Intelligent Systems*, 33(6):17–25, 2018.

[117] Alexander Prange, Mira Niemann, Antje Latendorf, Anika Steinert, and Daniel Sonntag. Multimodal speech-based dialogue for the mini-mental state examination. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, pages 1–8, New York, NY, USA, 2019. Association for Computing Machinery.

[118] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018.

[119] Dhanesh Ramachandram and Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108, 2017.

[120] Ekaterina Shutova, Douwe Kiela, and Jean Maillard. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 160–170, 2016.

[121] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[122] Alessandro Sordoni, Jing He, and Jian-Yun Nie. Modeling latent topic interactions using quantum interference for information retrieval. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 1197–1200, 2013.

[123] Victor J Stenger. *Timeless Reality: Symetry, Simplicity, and Multiple Universes*. Prometheus Books, 2009.

[124] Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. *arXiv preprint arXiv:1911.05544*, 2019.

[125] Chiheb Trabelsi, Olexa Bilaniuk, Dmitriy Serdyuk, Sandeep Subramanian, João Felipe Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher J. Pal. Deep complex networks. 2017.

[126] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.

[127] Yao-Hung Hubert Tsai, Martin Q Ma, Muqiao Yang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Interpretable multimodal routing for human multimodal language. *arXiv preprint arXiv:2004.14198*, 2020.

[128] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131, 1974.

[129] Amos Tversky and Daniel Kahneman. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, 90(4):293, 1983.

[130] Roope Uola, Kimmo Luoma, Tobias Moroder, and Teiko Heinosaari. Adaptive strategy for joint measurements. *Physical Review A*, 94(2):022109, 2016.

[131] Sagar Uprety, Shahram Dehdashti, Lauren Fell, Peter Bruza, and Dawei Song. Modelling dynamic interactions between relevance dimensions. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 35–42, 2019.

[132] Sagar Uprety, Dimitris Gkoumas, and Dawei Song. A survey of quantum theory inspired approaches to information retrieval. *ACM Computing Surveys (CSUR)*, 53(5):1–39, 2020.

[133] Sagar Uprety, Prayag Tiwari, Shahram Dehdashti, Lauren Fell, Dawei Song, Peter Bruza, and Massimo Melucci. Quantum-like structure in multidimensional relevance judgements. *Advances in Information Retrieval*, 12035:728, 2020.

[134] Cornelis Joost Van Rijsbergen. *The geometry of information retrieval*. Cambridge University Press, 2004.

[135] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

[136] Tomás Veloz, Xiaozhao Zhao, and Diederik Aerts. Measuring conceptual entanglement in collections of documents. In *International Symposium on Quantum Interaction*, pages 134–146. Springer, 2013.

[137] Valentin Vielzeuf, Stéphane Pateux, and Frédéric Jurie. Temporal multimodal fusion for video emotion classification in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 569–576, 2017.

[138] Pierpaolo Vivo, Mauricio P Pato, and Gleb Oshanin. Random pure states: Quantifying bipartite entanglement beyond the linear statistics. *Physical Review E*, 93(5):052106, 2016.

[139] Benyou Wang, Qiuchi Li, Massimo Melucci, and Dawei Song. Semantic hilbert space for text representation learning. In *The World Wide Web Conference*, pages 3293–3299, 2019.

[140] Benyou Wang, Peng Zhang, Jingfei Li, Dawei Song, Yuexian Hou, and Zhenguo Shang. Exploration of quantum interference in document relevance judgement discrepancy. *Entropy*, 18(4):144, 2016.

[141] Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P Xing. Select-additive learning: Improving generalization in multimodal sentiment analysis. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 949–954. IEEE, 2017.

[142] Jun Wang, Dawei Song, and Leszek Kaliciak. Tensor product of correlated text and visual features. In *In QI'10*. Citeseer, 2010.

[143] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7216–7223, 2019.

[144] Jason Weston, Emily Dinan, and Alexander H Miller. Retrieve and refine: Improved sequence generation models for dialogue. *arXiv preprint arXiv:1808.04776*, 2018.

[145] Dominic Widdows. Orthogonal negation in vector spaces for modelling word-meanings and document retrieval. In *Proceedings of the 41st annual meeting of the association for computational linguistics*, pages 136–143, 2003.

[146] Dominic Widdows and Dominic Widdows. *Geometry and meaning*, volume 773. CSLI publications Stanford, 2004.

[147] Scott Wisdom, Thomas Powers, John Hershey, Jonathan Le Roux, and Les Atlas. Full-capacity unitary recurrent neural networks. In *Advances in neural information processing systems*, pages 4880–4888, 2016.

[148] Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. Proactive human-machine conversation with explicit conversation goal. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3794–3804. Association for Computational Linguistics, 2019.

[149] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016.

[150] Jiahong Yuan and Mark Liberman. Speaker identification on the scotus corpus. *Journal of the Acoustical Society of America*, 123(5):3878, 2008.

[151] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017.

[152] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. *arXiv preprint arXiv:1802.00927*, 2018.

[153] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, volume 2018, page 5642. NIH Public Access, 2018.

[154] Amir Zadeh, Chengfeng Mao, Kelly Shi, Yiwei Zhang, Paul Pu Liang, Soujanya Poria, and Louis-Philippe Morency. Factorized multimodal transformer for multimodal sequential learning. *arXiv preprint arXiv:1911.09826*, 2019.

[155] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016.

[156] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016.

[157] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018.

[158] Anton Zeilinger. Experiment and the foundations of quantum physics. In *More Things in Heaven and Earth*, pages 482–498. Springer, 1999.

[159] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 2020.

[160] Chenguang Zhang, Yuexian Hou, and Dawei Song. Quantum observation scheme universally identifying causalities from correlations. *Physical Review A*, 101(6):062103, 2020.

[161] Peng Zhang, Jingfei Li, Benyou Wang, Xiaozhao Zhao, Dawei Song, Yuexian Hou, and Massimo Melucci. A quantum query expansion approach for session search. *Entropy*, 18(4):146, 2016.

[162] Peng Zhang, Jiabin Niu, Zhan Su, Benyou Wang, Liqun Ma, and Dawei Song. End-to-end quantum-like language models with application to question answering. 2018.

[163] Yazhou Zhang, Dawei Song, Xiang Li, Peng Zhang, Panpan Wang, Lu Rong, Guangliang Yu, and Bo Wang. A quantum-like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis. *Information Fusion*, 2020.

[164] Yazhou Zhang, Dawei Song, Peng Zhang, Panpan Wang, Jingfei Li, Xiang Li, and Benyou Wang. A quantum-inspired multimodal sentiment analysis framework. *Theoretical Computer Science*, 752:21–40, 2018.

[165] Yuanyuan Zhang, Zi-Rui Wang, and Jun Du. Deep fusion: An attention guided factorized bilinear pooling for audio-video emotion recognition. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[166] Guido Zuccon and Leif Azzopardi. Using the quantum probability ranking principle to rank interdependent documents. In *European Conference on Information Retrieval*, pages 357–369. Springer, 2010.